

2010年中国信息技术应用 学术研讨会论文集

—— ◎ 主编 林志远 副主编 张 月 岳冰冰 ——



2010 年中国信息技术应用 学术研讨会论文集

主 编 林志远

副主编 张 月 岳冰冰

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书是中国信息产业商会信息技术应用分会的重要学术载体,共收集了43篇学术论文,内容涉及信号与数据处理、电路与系统、计算机技术与应用、网络理论与技术、信息系统集成技术等,反映了近年来国内信息技术研究与应用的新的学术成果,内容丰富,涉及面广,专业性强,是国内研究信息技术及其应用的重要参考文献。

本书适合于高等院校信息科学技术领域的师生,IT类科研机构的专家、学者,信息科技相关政府主管部门的管理者以及企事业单位技术信息技术应用的实践工作者学习、借鉴和参考。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

2010年中国信息技术应用学术研讨会论文集/林志远主编. —北京:电子工业出版社,2010.6
ISBN 978-7-121-11028-3

I. ①2… II. ①林… III. ①计算机应用—学术会议—文集 IV. ①TP39-53

中国版本图书馆CIP数据核字(2010)第104157号

责任编辑:董亚峰

印 刷: 北京季蜂印刷有限公司

装 订: 电子工业出版社

北京市海淀区万寿路173信箱 邮编 100036

开 本: 787×1 092 1/16 印张: 19.5 字数: 610千字

印 次: 2010年6月第1次印刷

定 价: 88.00元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888。

质量投诉请发邮件至zltts@phei.com.cn,盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线:(010)88258888。

前 言

21 世纪以来，信息技术不断创新，信息产业持续发展，信息技术应用不断深入，信息化成为全球经济社会发展的显著特征。广泛应用、高度渗透的信息技术正孕育着新的重大突破，信息技术应用也成为当今世界发展的大趋势，成为推动经济社会变革的重要力量。

信息技术在我国国民经济和社会各领域的应用效果日渐显著。农业信息服务体系不断完善；应用信息技术改造传统产业不断取得新的进展，能源、交通运输、冶金、机械和化工等行业的信息化水平逐步提高；传统服务业转型步伐加快，信息服务业蓬勃兴起；金融信息化推进了金融服务创新，现代化金融服务体系初步形成；电子商务发展势头良好，科技、教育、文化、医疗卫生、社会保障、环境保护等领域信息化步伐明显加快。

大力推进信息技术应用，是覆盖我国现代化建设全局的战略举措，是贯彻落实科学发展观、全面建设小康社会、构建社会主义和谐社会和建设创新型国家的迫切需要和必然选择。但是，我国信息技术应用水平不高，与先进国家相比存在较大差距。在整体上，应用水平落后于实际需求，信息技术的潜能尚未得到充分挖掘，在部分领域和地区应用效果不够明显；信息技术自主创新能力不足，核心技术和关键装备主要依赖进口；信息安全问题仍比较突出；数字鸿沟有所扩大；国内不同地区、不同领域、不同群体的信息技术应用水平和网络普及程度很不平衡，城乡、区域和行业的差距有扩大趋势。

为贯彻落实《2006—2020 年国家信息化发展战略》，及时反映和交流信息技术在经济社会各领域广泛应用的最新学术研究成果，中国信息产业商会信息技术应用分会开展了“2010 年中国信息技术应用学术研讨会”征文活动。征文活动共收到论文 300 余篇，经过严格评审，精选出优秀论文 43 篇，内容涵盖信息技术发展与应用的诸多领域，以促进我国信息技术应用的学术交流与借鉴。

限于时间、条件和水平，本论文集难免存在疏漏，欢迎读者批评指正。

中国信息产业商会信息技术应用分会

二〇一〇年六月二十六日

目 录

第 1 部分 计算机技术与应用

智能故障诊断技术的研究及应用	李爱民 (3)
构件技术在食品安全可追溯系统中的应用	李骏平 王燕兴 (9)
基于 FPGA 的 UART 的设计与仿真	李 琦 董利民 吴武臣 侯立刚 (15)
运动图的方法计算合成双人运动	刘 宁 (22)
全过程计算机辅助古代塔类建筑动画自动生成	刘射彪 梁天柱 (29)
高校办公室知识管理维度分析	陈学东 刘文娟 (39)
Return to the Origin of Architectural Design: Based on the Research and Application of Sketchup in the Teaching of Architectural Design	Tang Hong Wang Jinyu (44)
自主水下机器人能源系统设计	滕学志 魏志强 殷 波 董 艳 (52)
电子设计综合实验教学改革研究	王建新 李 莉 高献伟 路而红 (58)
研究性教学模式下电子信息类课程教学改革探索	王建新 李秀滢 周玉坤 陈汉林 (62)
动画自动生成中基于分镜头的摄像机规划	王巍峰 (66)
机场柜台资源共享解决方案—中国民航离港多主机共用平台系统	王欣明 高 新 (74)
Research on Pork Safety Traceability System Based On RFID	Tong Xinshun Wu Yi (82)
基于 GTechnology 的输电网 WebGIS 的 设计与实现	徐雪荣 郭世界 王晓辉 (89)
虚拟现实中的力/触觉建模技术	徐玉彬 刘玉庆 朱秀庆 (95)
古建动画自动生成系统中构件位置计算	尹梅芳 未官瑾 (103)
IT 应用与经管教学融合的探讨	张 耀 崔锦荣 (113)

第 2 部分 数字信号处理

Short-Term Load Forecasting Based on Dynamic Recurrent Fuzzy Neural Network	Ge Chao Zhang Jingchun Sun Yanbin Sun Liying (119)
基于信息熵的模糊聚类信誉评价体系	宫尚宝 郭玉翠 (126)
平面紧凑型双通带滤波器设计	刘艳萍 (133)
一种基于用户背景知识的文本聚类方法	沈志辉 袁再江 (138)
对一类非线性系统基于三角反馈的 Hopf 分岔控制	童 炜 万里红 (143)
基于属性相似性计算的空间关联规则提取技术研究	王海涛 (150)

最快完成车辆路径问题的改进蚁群算法	闻思源	魏红翠	(155)
基于训练文本特征扩展的中文短文本分类研究	闫 涛	王细薇	樊战伟 (162)
基于 Jacobi 旋转的稀疏矩阵对角优势强化方法	银福康	宋君强	吴建平 (169)
服从二维指数分布的非独立随机变量的 线性组合的分布	郭云飞	尹 哲	(177)
一种 PDF 信息提取与表格重现的算法	张 伯	陈 彩	(182)
An Estimate Method for the Maximum Maintenance Time of Normal Distribution Items	Jin Xing	Peng Bo	Lu Hai (189)

第 3 部分 通信理论与技术

一种认知无线电频谱感知与接入的联合设计方案	丛 容	吴迎笑	(197)
多用户检测算法及其 simulink 仿真研究	任大山	龙 昕	杨明华 (204)
On Design and Simulation of Electrostatic Sensor Used for Measuring Gas-Solid Two-phase Flow		Yu Zhigen	(210)
协作分集中的移动中继动态选择和切换策略研究	张 鑫	谢显中	雷维嘉 (218)

第 4 部分 网络理论与技术

基于无线传感器网络的危险货物运输系统	陈 晨	裴庆祺	庞辽军	张素兵	范科峰 (235)
军用 ASON 融合建网的应用可行性解析	任志宏	谢永强		赵广松	(241)
IPTV 业务在现有宽带网络中的实现				赵 怡	(249)

第 5 部分 系统集成技术

软件项目计划与跟踪	刘卫宏	焦彦平	(259)
基于企业营销渠道管理的信息服务		宋 瑾	(266)
基于 DEA 方法的商业银行信息化效率评价	王江涛	邱 月	(271)

第 6 部分 信息安全

浅谈信息安全中的加密技术	李左伦	杜春梅	郭 鑫	(281)
一种改进木材细胞图像分形特征提取方法	任洪娥	高 莹	董本志	(285)
一种基于 TPCM 可信移动存储设备的发行与认证	阮富生	王 冠	刘智君	王 博 (289)
基于 P2P 文件共享系统的抗攻击信任管理模型	吴 旭	郭玉翠	宫尚宝	(295)
一种密码算法数据模型的设计与实现	周修义	何新民	徐莉伟	(300)

第 1 部分

计算机技术与应用

智能故障诊断技术的研究及应用

李爱民

(海军工程大学 船舶与动力学院, 湖北 武汉 430033)

摘 要: 智能故障诊断技术的关键是基于数据预处理的工况状态识别。本文首先介绍了贝叶斯故障诊断、模糊故障诊断、基于粗糙集理论的故障诊断、神经网络故障诊断、基于成分分析的故障诊断和专家系统故障诊断的原理和方法; 然后比较了各种诊断方法的特点, 分析了各种诊断方法的应用; 最后提出了智能故障诊断技术的发展趋势。

关键词: 故障诊断; 状态识别; 人工智能

Research and Application of Intelligent Fault Diagnosis Methods

LI Ai-min

(College of Naval Architecture and Power, Naval University of Engineering,
Hubei Wuhan 430033, China)

Abstract: The key of intelligent fault diagnosis technology is state recognition based on data processing methods. Firstly, the principles and methods of Bayes fault diagnosis, fuzzy fault diagnosis, rough set fault diagnosis, neural networks fault diagnosis, principal components analysis and independent component analysis fault diagnosis and expert system fault diagnosis were mainly introduced. Then, the characteristics of the methods were compared, at the same time the applications of the methods were analyzed. Finally, the development trends of intelligent fault diagnosis technology were advanced.

Keywords: fault diagnosis; state recognition; artificial intelligence

1 引言

故障诊断是一门发展中的新兴学科, 任务是根据机械设备的运行信息来识别设备的相关状态, 实质就是状态识别。主要环节包括信号检测、特征分析与提取、工况状态识别和故障诊断。目前, 机械设备的故障诊断已成为保证生产线安全运行、提高产品质量的重要手段和关键技术。由于机械设备自身机构、运行过程和环境的复杂性, 其运行特征参数与状态之间并不全是一一对应的关系, 致使故障诊断方法非常复杂。人工智能和计算机技术的发展和应用于智能故障诊断开辟了新途径。

智能诊断技术是以人工智能及计算机技术为基础, 判断系统故障的所属类别和严重程度,

其关键是基于数据预处理的工况状态识别方法。主要有贝叶斯分类法、模糊分类法、成分分析法、基于粗糙集理论的诊断方法、神经网络诊断法和专家系统诊断等。下面结合数据预处理方法对故障诊断技术的原理和应用进行分析和介绍。

2 基于人工智能的故障诊断方法

1) 贝叶斯分类法

贝叶斯分类法是以概率统计为基础来描述工况状态的变化。在机械系统中大量的事件都是随机的，但它按照某种统计规律而发生变化，事件出现的概率在很多情况下是可以估计的。这种根据先验知识对工况状态出现的概率做出的估计，称为先验概率。由于状态是随机变量，故状态空间记为 $\Omega_j = (\omega_1, L, \omega_i, L, \omega_m)$ ，其中 $\omega_i (i=1, 2, L, m)$ 是状态空间中的一个模式点。设第 i 类的先验概率为 $P(\omega_i)$ ，并有 $\sum_{i=1}^m P(\omega_i) = 1$ ，类条件概率密度为 $p(x/\omega_i)$ ，根据 Bayes 公式有：

$$P(\omega_i / x) = \frac{p(x/\omega_i)P(\omega_i)}{\sum_{j=1}^m p(x/\omega_j)P(\omega_j)} \tag{1}$$

其中 $P(\omega_i / x)$ 表示已知样本条件下 ω_i 出现的概率，称为后验概率。通过（1）式把观测值 x 的先验概率转化为后验概率。

贝叶斯判别法的决策规则主要有最小错误率规则、最小平均损失规则以及考虑 $P(\omega_i)$ 变化的最小最大规则，应用中应根据实际情况选择合适的决策规则设计状态分类器。

基于贝叶斯诊断的特点在于利用先验概率为诊断服务，这就依赖于先验知识的积累。

2) 模糊诊断方法^[1]

模糊故障诊断是通过研究故障与征兆之间的模糊关系来判别设备的运行状态。系统中可能发生的故障用状态论域表示为 $\Omega = \{\omega_1, \omega_2, L, \omega_m\}$ ，其中 m 为故障的种数；与故障有关的各种特征用征兆论域表示为 $K = \{K_1, K_2, L, K_n\}$ ， n 为征兆的种数。论域中的元素均有各自的隶属函数，如 ω_i 的隶属函数为 $\mu_{\omega_i} (i=1, 2, L, m)$ ， K_j 的隶属函数为 $\mu_{K_j} (j=1, 2, L, n)$ 。其矢量形式为 $A = [\mu_{\omega_1}, \mu_{\omega_2}, L, \mu_{\omega_m}]^T$ ， $B = [\mu_{K_1}, \mu_{K_2}, L, \mu_{K_n}]^T$ ，称 A 为故障模糊矢量，是故障在状态论域 Ω 上的表现； B 为特征模糊矢量，是故障在征兆论域 K 上的表现。故障的模糊诊断，可以认为是状态论域 Ω 与征兆论域 K 之间的模糊矩阵运算。模糊关系方程为

$$A = R * B \tag{2}$$

式中 $R = [r_{ij}]$ ，称为模糊关系矩阵， $r_{ij} \in [0, 1]$ 表示第 j 种征兆 K_j 对第 i 种故障 ω_i 的隶属度；“*”为广义模糊逻辑算子，表示不同的逻辑运算。

通过（2）式确定待检状态的模糊矢量，然后依据模糊诊断准则大致确定故障。常用的诊断准则有最大隶属度准则、择近准则和模糊聚类准则等。

选择合适的隶属函数和确定模糊关系矩阵是模糊诊断中的关键技术，需要参考大量故障诊断的经验和实验测试。

3) 基于粗糙集理论的故障诊断

粗糙集理论是一种用于处理不完整、不精确知识的数学方法，该理论不需要关于数据的任何初始或附加信息，直接对不完整、不精确数据进行分析处理，发现数据之间的关系，提取有用特征，得到简明扼要的知识表达形式^[2]。基于粗糙集理论的故障诊断过程为^[3]：

- ① 获取原始数据样本，确定初始决策表；
- ② 对初始决策表进行离散化处理，得到离散化决策表；
- ③ 对决策表进行约简，得到基于分类模式的决策表。

粗糙集的处理对象为离散型数据，故障检测所得量的属性为连续型数据，因此必须先进行离散化处理，离散结果将影响导出规则的统计特性。因此连续属性的离散化方法是基于粗糙集理论故障诊断的关键。

4) 基于神经网络的故障诊断

人工神经网络（ANN）是由大量的简单非线性处理单元——人工神经元高度错综复杂连接而成的网状系统。人工神经元的特性、连接拓扑结构和学习规则是该网络的三要素。该网络的自学习、自组织、联想记忆及容错等功能能较好地处理不确定的、矛盾的、甚至错误的信息，在故障诊断领域受到广泛关注。它通过对故障实例及诊断经验的训练和学习，用分布在神经网络中的连接权值来表达所学习的故障诊断知识，具有对故障联想记忆、模糊匹配和相似归纳等能力^[4]，可实现故障与征兆之间的非线性对应关系。

用于故障诊断的有多层前向网络、径向基函数网络、模糊神经网络、无监督神经网络等。

5) 基于主成分分析及独立成分分析的故障诊断

主成分分析（PCA）是多元统计分析中除去数据相关性、进行特征降维的统计分析方法。基本思路是选择相对较少的线性无关的新变量来概括原有数据的主要特征，将高维数据投影到低维空间。主成分可以通过求解样本协方差矩阵的特征值和特征向量获得，相关算法这里不再赘述。

独立成分分析（ICA）是从观测到的多个独立信号的混合信号中分离出源信号，实现信号的盲源分离。思路是通过一个反混合矩阵 W 对测试信号进行变换以获得相应的独立源信号。反混合矩阵 W 要通过自组织学习的方法获得。建立以 W 为变量的目标函数 $L(W)$ ，然后通过优化目标函数来估计 W 。ICA 的目标函数有统计独立性函数、极大似然函数、最大熵函数等。

PCA 和 ICA 是消除冗余信息、提高诊断效率的有效的方法。两者的区别在于：PCA 是通过卡亨南—洛维变换（K-L 变换）以获得互不相关的特征，而 ICA 是通过线性变换将原始信号分解为相互统计独立的分量。因此可以把 ICA 看作 PCA 的高阶扩展。基于 PCA 或 ICA 的故障诊断过程如图 1 所示^[1]。

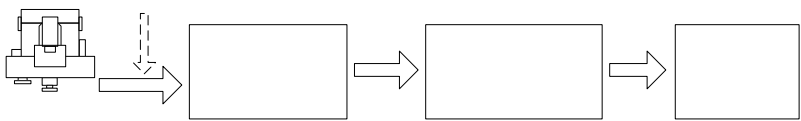


图 1 基于 PCA 或 ICA 的故障诊断过程

6) 基于专家系统的故障诊断

专家系统是应用人类专家的知识和推理方法求解复杂问题的一种人工智能程序^[1]。专家系统中，专家的知识在分离的独立知识库中进行描述，每个知识单元描述一个具体情况，通过

推理机制可以对不同的处理对象从知识库中选择不同的知识单元构成不同的求解序列, 完成指定任务。专家系统由知识库、数据库、推理机、解释程序和知识获取程序构成。图 2 为一个实用的专家系统框图^[1], 除前述的 5 个基本模块(图中有阴影的方块)外, 还有许多中间环节模块。

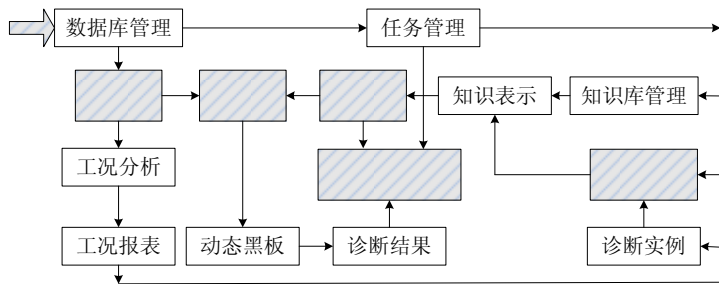


图2 实用专家系统框图

基于专家系统的故障诊断策略大体上分为两种^[1]：一种是“基于知识”的诊断策略，采用“自上而下”的思路，把来自不同的机器对某种故障获得的“知识”构成知识空间，对某一实际机器的异常工况进行诊断。另一种是“基于行为”的诊断策略，采用“自下而上”的思路，从某一台机器的实际运行状态出发，从其工况状态的变化判断其故障属性。“基于知识”的诊断策略要解决的基本问题是知识获取，因为这种策略导致故障样本的可比性差，误判率大。另一个问题是机器的自学习能力差，人工干预多。“基于行为”的诊断策略的核心思路是：一个诊断系统应能够在其运行过程中，不断地提高自身的智能化水平，即诊断系统应当具有智力进化的功能。

3 各种智能诊断方法的比较和应用

1) 各种诊断方法的比较

① 贝叶斯诊断是以概率统计为基础的,模糊集和概率统计方法是处理不确定信息的常用方法,但这些方法需要数据的附加信息或先验知识,如隶属函数和概率分布等,这些信息有时不易得到^[5]。

② 粗糙集方法仅利用数据本身提供的信息, 无需任何先验知识, 是以不可分辨关系为基础, 侧重分类; 模糊集则基于元素对集合隶属度的不同, 强调集合本身的含混性。从粗糙集的观点看, 粗糙集不能清晰定义的原因是缺乏足够的论域知识, 但可以用清晰集合逼近^[5]。

③ 神经网络具有以任意精度逼近连续非线性函数的能力;网络并行工作,速度极快;信息存储于网络的各个权值之中,某些单元障碍并不影响信息的完整,具有鲁棒性。不足之处在于网络权值的物理含义不明确,忽视了领域专家的经验知识等。因此在故障诊断中常将神经网络与模糊诊断、专家系统相结合,充分发挥神经网络独特的优势。与贝叶斯诊断相比,神经网络分类器的有效性优于贝叶斯分类器^[6];与粗糙集诊断法相比,它既可以处理连续属性的模拟量,也可以处理混沌的、不完全的、模糊的信息。

④ 专家系统的优势在于集专家知识和各种智能诊断手段于一身,不足之处在于知识获取困难、知识库更新能力差。对于新型设备往往无从获得诊断知识;对于规范化差的设备,由

于工作特性与规范化设备相比差异太大,知识获取也很困难。

由上可见:每种诊断方法都有优缺点。故障诊断中常将这些方法结合起来使用,发挥各自的优势,提高诊断的准确性和效率。

2) 智能故障诊断技术的应用

智能故障诊断技术在工程领域的应用非常广泛。例如 ICA 在转子系统、齿轮箱故障诊断中的应用^[7, 8],改变了传统的以降噪为主的故障信息增强思想,为微弱故障的有效诊断提供了技术手段。除单一诊断方法外,实际应用中更多的是将各种诊断方法综合在一起,实现故障诊断方法的优势互补。例如将贝叶斯方法与粗糙集相结合的变压器综合故障诊断、汽轮机振动故障诊断^[9, 10],这种结合具有处理信息缺失多的能力和容错特性,克服了粗糙集刚性推理的弱点,其性能明显优于单独使用贝叶斯网络分类器或粗糙集的方法;将粗糙集与神经网络结合的故障诊断^[11, 12, 13],利用粗糙集的约简功能消除冗余信息,再与神经网络相结合,提高诊断的准确性和效率。专家系统的成功应用有水轮发电机组故障诊断模糊专家系统^[14]、导弹发射车液压系统的诊断专家系统^[15]、基于粗糙集理论的变压器故障诊断专家系统^[16]、基于神经网络和专家系统的电传操纵系统故障^[17]等,它们充分地将多种知识获取方法应用于专家系统,解决了故障诊断中不确定性问题的求解,满足了故障诊断的有效性和实用性要求。

4 结束语

随着人工智能和计算机技术的发展,机械设备的故障诊断技术也有了很大飞跃。目前智能故障诊断研究的主要内容有:①运用新的智能方法完善故障诊断内容,如将演化算法、分形理论等应用到故障诊断中,提高故障诊断的能力;②研究复合式故障诊断方法,如将模糊理论、粗糙集理论、神经网络及信息融合技术等运用于故障诊断,突出各种方法的优势;③研究面向 Web 的多代理故障诊断系统,由于面向 Web 的多代理故障诊断能实现 Internet 环境下的远程诊断,消除信息“孤岛”效应,实现资源共享,提高诊断精度,因此这方面的研究也将成为故障诊断研究的一个重要内容。

参考文献

- [1] 钟秉林,黄仁.机械故障诊断学(第3版)[M].北京:机械工业出版社,2006.
- [2] 曲晓慧,郑利庆,安钢.基于粗糙集理论的机械状态监测与故障诊断[J].测试技术学报,2009,23(2):178-182.
- [3] 高赞,侯媛彬,朱华.基于粗糙集理论的系统建模方法[J].长安大学学报,2005,25(2):98-101.
- [4] 何勇,李增芳.智能化故障诊断技术的研究与应用[J].浙江大学学报,2003,29(2):119-124.
- [5] 纪滨.粗糙集理论及进展的研究[J].计算机技术与发展,2007,17(3):69-72.
- [6] 胡芑庆,温熙森.船用汽轮机减速箱运行状态检测方法[J].国防科技大学学报,1997,19(6):36-41.
- [7] 黄晋英,毕世华,潘宏侠等.独立分量分析在齿轮箱故障诊断中的应用[J].振动、测试与诊断,2008,28(2):35-39.
- [8] 郝志华,张一杨,刘岩.独立成分分析在转子故障诊断中的应用[J].汽轮机技术,2007(2):29-32.
- [9] 董泽,张楠,韩璞.粗糙集理论和贝叶斯网络在汽轮机振动故障诊断中的应用[J].汽轮机技术,2008(03):15-19.

- [10] 朱永利, 吴立增, 李雪玉. 贝叶斯分类器与粗糙集相结合的变压器综合故障诊断[J]. 中国电机工程学报, 2005, 25(10):17-21.
- [11] 雷霆, 代传龙, 王厚军等. 粗糙集-神经网络集成的 WSN 节点故障诊断[J]. 电子科技大学学报, 2008, 37(4):565-568.
- [12] 谢振华, 商琳, 李宁等. 粗糙集在神经网络中应用技术研究[J]. 计算机应用研究, 2004,21(9):71-74.
- [13] 范兴铎, 盛颂恩. 基于粗糙集-神经网络的智能混合压缩机故障诊断系统的研究[J]. 压缩机技术, 2007, (2):1-3.
- [14] 刘晓波, 黄其柏. 水轮发电机组故障诊断模糊专家系统研究[J]. 华中科技大学学报, 2006,34(1):43-47.
- [15] 周汝胜, 焦宗夏, 王少萍等. 基于专家系统的导弹发射车液压系统故障诊断[J]. 航空学报, 2008, 29(1):59-63.
- [16] 项新建. 基于粗糙集理论的变压器故障诊断专家系统研究[J]. 仪器仪表学报, 2005, 26(1):22-26.
- [17] 徐荣红, 孙金标. 基于神经网络和专家系统的电传操纵系统故障诊断[J]. 航空学报, 2005, 26(2):47-51.

作者简介

李爱民, 男, 1978 年 2 月生, 硕士, 讲师, 主要研究方向为机械设备状态监测与故障诊断。

构件技术在食品安全可追溯系统中的应用

李骏平 王燕兴

(北京工业大学计算机学院, 北京, 100124)

摘要: 在食品安全可追溯系统软件项目的设计过程中, 采用基于领域工程的方法, 为食品安全可追溯领域确定边界和制定需求, 通过领域分析在已有经验基础上建立领域模型, 确定领域的共同特征, 定义通用的领域静态和动态结构模型, 最终开发出可复用构件。把食品安全可追溯系统开发问题转化为对领域构件的设计、使用问题, 增强了系统的灵活性和通用性。这将使食品安全可追溯系统的开发升级更加快捷、方便, 追溯系统也将得到更好的推广。

关键字: 食品安全追溯; 领域分析; 软件复用; 构件技术

The Application of Component Technology in the Food Safety Traceability System

LI Jun-ping WANG Yan-xing

(Colleague of Computer Science, Beijing University of Technology, Beijing 100124)

Abstract: In the design of the Food Safety Traceability System, domain engineering is used to confirm the boundaries and get the requirements. It uses domain analysis which is based on the existing experience in the domain model to identify common features of domain, define the common static and dynamic structural model of domain, eventually obtain the reusable component. The development of food safety system can be translated into the design and using of the domain component, that can enhance the flexibility and versatility of the system. This will be faster and more convenient to develop and update the system, and these traceability system will also be better to popularize.

Keywords: Food Safety Traceability; Domain analysis; Software reuse technology; Component technology

1 背景

近年来, 各国食品安全领域屡屡出现问题, 从国外的疯牛病、口蹄疫到我国的注水肉、问题奶粉、苏丹红事件、三聚氰胺事件等, 引起了世界的广泛关注。如何保障食品安全已成为消费者和经营者共同关注的课题, 而且已成为影响我国农业和食品产业国际竞争力的重要因素^[1]。国家积极应对出现的各种问题, 进行了食品可追溯系统的开发研究, 制定了一些相关

的标准和指南, 在一些有条件适合地方和企业初步建立了部分食品可追溯制度与法规, 并建立了一批可追溯食品和可追溯企业, 形成了一系列的追溯子系统, 由我国物品编码中心和中国食品工业协会合办, 构建了商品条码食品安全追溯平台, 让消费者了解符合卫生安全的生产和流通过程, 提高消费者放心程度。

目前, 各种不同形式的追溯工程普遍存在缺乏理论指导和统一规划、建设目标不清晰、工程设计不合理、缺少技术标准体系等问题, 从根本上解决只停留在信息的追溯上而没有通过安全流通控制理论来保证传递信息不被篡改, 传递信息的可信问题。

2 食品安全可追溯系统应用现状

目前我国, 谷物、水果、肉类、禽蛋和水产品等主要食品产量居世界第一位, 为了保证广大人民群众的食品安全, 以及排除我国食品的出口面对他国食品跟踪与追溯法律法规的限制, 因此在我国建立食品追溯的工作将对食品行业的发展产生巨大的影响。但当前我国在整个食品生产过程中应用自动追溯系统的实例仍寥寥无几, 国内食品行业追溯目前还主要仅仅是在零售结算环节, 远未在食品供应链的全过程应用, 全程可跟踪供应链尚未形成。

“食品质量安全追溯系统”是一个能够连接生产、检验、监管和消费各个环节, 让消费者了解符合卫生安全的生产和流通过程, 提高消费者放心程度的信息管理系统。该系统提供了“从农田到餐桌”的追溯模式^[1], 提取了生产、加工、流通、消费等供应链环节消费者关心的公共追溯要素, 建立了食品安全信息数据库, 一旦发现问题, 能够根据溯源进行有效的控制和召回, 从源头上保障消费者的合法权益。

3 基于构件的食品安全可追溯系统

构件(Component)是指应用系统中可以明确辨识的构成成分。在计算机百科全书中, 软件构件被定义为软件系统中具有相对独立功能、可以明确辨识、接口由契约指定、和语境有明显依赖关系、可独立部署、且多由第三方提供的可组装软件实体^[2]。

1) 构件技术的特点

构件的一个主要特性就是封装性, 其内部封装所能提供的功能。对于构件使用者而言, 完全不必知道构件的实际内容, 只需把它作为黑盒子通过接口调用它即可。因此构件的接口决定了构件与外界的联系, 甚至在某种程度上可以认为构件的接口决定了构件本身, 整个接口的设计是构件开发过程中尤为重要的一点^[3]。在不同层次上, 构件均可以将底层的多个逻辑组合成高层次上的粒度更大的新构件, 甚至直接封装到一个系统, 使模块的重用从代码级、对象级、架构级到系统级都可能实现, 从而使软件像硬件一样, 能任人装配定制而成的梦想得以实现。

2) 领域工程提取可复用的构件

领域工程是软件开发者用于为一类相似或相近系统或应用软件定义范围、指定结构和创建可复用资产的基于复用的过程和实践, 它包括界定、分类和创建可复用构件的所有活动。而“领域”是指一组具有公共属性的系统^[4]。

在复用实施的过程中, 复用基础设施是不断演化的。在开发新系统时, 开发者从复用基

基础设施中获得可复用信息，将它们集成到新系统中去。在系统开发过程中可能产生对于信息描述、分类方式等方面的反馈意见，并以此调整复用基础设施。当系统开发完成，复用基础设施就需要根据当前现有系统进行演化，并将本领域中系统的共同特征提取出来。复用基础设施的演化，是可以并且需要随着领域中系统的开发持续进行的^[5]。

众多追溯领域中的系统，在功能上有类似之处，比如都有生产或养殖阶段，有加工处理或屠宰阶段，有批发销售阶段等等，这些大的子系统或子模块下面的业务也在不同的粒度上有着相同或相似的功能与处理流程。因而利用构件技术提高复用率，大大简化开发过程，提高开发效率。

3) 食品安全可追溯系统构件设计

食品安全可追溯系统在食品追溯领域具有良好的适应性和灵活性，通过大量构件的使用，很好地解决了流通方式的差异对软件系统的影响，实现以较小的代价适应制度与业务的变化。除此之外还能有效地防止信息被篡改，实现物流、信息流和管理流的三流一体化体系结构，确实、可靠的保证每一件食品流通过程中的每一个环节均可查询到最为准确的信息，有效地避免了信息的干扰和篡改，实现食品安全预警机制。而基于可信计算的安全追溯系统，其可复用性和可扩展性的要求比一般的系统更高。而为了更好地满足这些质量需求，构件技术正是一个好的选择。

食品安全可追溯系统的业务逻辑功能划分如图 1 所示。其核心是养殖、屠宰、批发、销售、追溯功能。

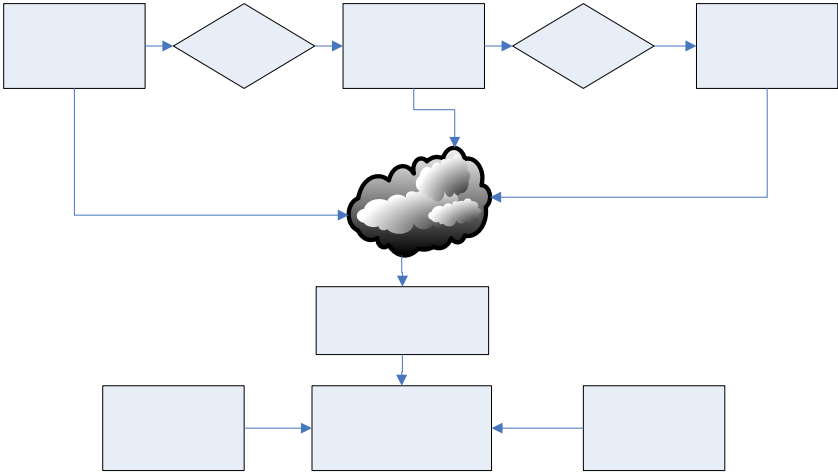


图 1 食品安全可追溯系统功能图

领域工程的过程包括领域划分、领域分析、体系结构的开发、资产生产。在本系统的开发过程为：首先，从业务角度分析整个系统，对其按功能模块划分，如上图的业务流程模块，即进行领域划分给出业务边界。再在此基础上进行领域分析，进行知识获取，将获取的知识组织到领域模型中，根据现有系统、标准规范等验证领域模型的准确性和一致性，维护领域模型。最后根据领域模型提取或开发相应的可复用构件。最后得出本系统的系统构件划分如图 2 所示。

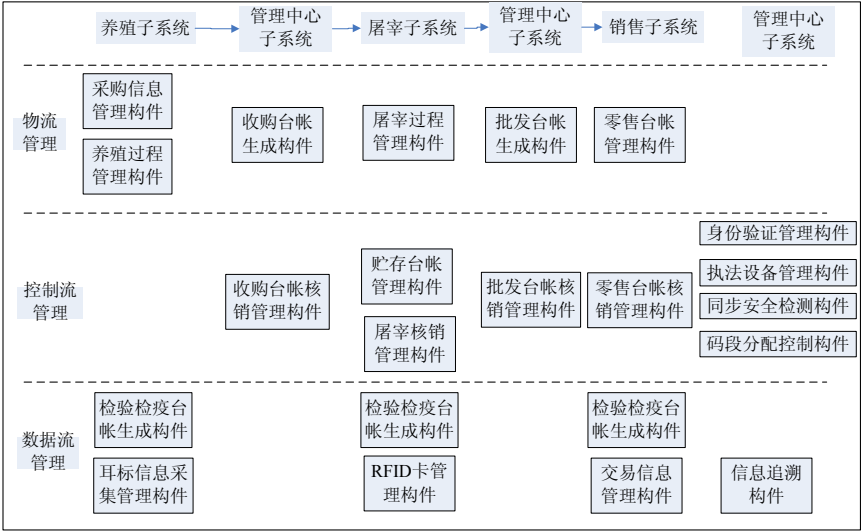


图 2 食品安全可追溯系统构件划分图

以下为子系统内的构件提取实例。在猪肉食品安全可追溯系统中，以屠宰阶段为例，如图 3 所示，在屠宰厂里，包括活猪的进厂、活猪屠宰、冷藏、批发出厂的 4 大流程。

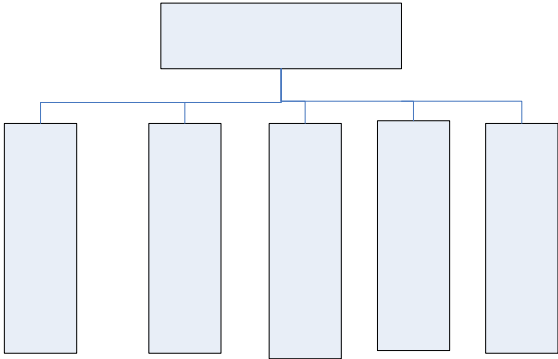


图 3 屠宰子系统构件划分图

在活猪进厂处理环节的构件是 RFID 卡管理构件，此构件的功能是记录屠宰阶段的追溯关键信息屠宰号和记录活猪的进厂相关信息。在此构件中，首先记录各批次的活猪进场信息，并给出对应的唯一的批号（由日期和批次构成），并负责对在进入屠宰间之前的活猪预处理，如清洗、去耳标、个体编号等操作的进行记录。在传统的食品追溯系统中，这些都是属于屠宰阶段的功能模块之中，只在屠宰场内部系统中使用，或者是与此有直接业务关系的上下游企业分别订制的系统，因此不具有灵活的适应性和可复用性。遇到有些养殖场不具有活猪佩戴耳标的生产条件时，这样的系统的屠宰阶段就不能接收这些活猪，从而使得该系统并不适合该地区或部门使用。而与此类似的适应性问题在国内庞大的食品消费市场中是存在很多且差距巨大的。

因此，本系统在 RFID 卡管理构件中，包含多个子构件，其中有关耳标处理(记录耳标信息)功能的构件提供两个接口。一个提供针对有耳标的活猪处理的服务，在摘除每头活猪的耳标时，系统会自动记录耳标信息，并生成耳标信息追溯记录。另一个提供针对无耳标的活猪处理的服务，在摘除每头活猪的耳标时，系统会自动记录耳标信息，并生成耳标信息追溯记录。

标时记录其耳标号,并同时给出屠宰号(包含耳标号);另一个则提供针对无耳标的活猪处理的服务,中间省去了耳标信息记录,但为了保持可扩展性和灵活性,在构件内部由系统给定了一个虚拟的耳标号。这样这两个构件都具有统一的对外接口,在组装不同的系统时,只需按照需求选择符合相应条件的构件即可。如果还有更个性化的需求,则各系统可自行编写相应接口的个性实现,并组装到系统中,当然前提是符合领域模型给定的统一接口规范与说明。

4) 应用构件技术对食品可追溯领域的影响

由前言可知,在国内食品追溯系统是不健全的,很多地方都没有建立追溯体系,特别是中小城市。而从国家发展的指导思想中可以看出,全国范围内建立食品追溯系统是发展的大势,而应用构件技术则会使这一进程加速,并从一开始就站在了展望未来的制高点,不仅保证了产品的质量,而且大大缩短了开发周期,节约了开发成本和实施成本。

构件化的追溯系统将会调整食品生产销售产业链,并将改变食品追溯的运营模式。传统追溯系统的运营模式是一种落后的、单干的商业模式,那就是养殖、加工、追溯由一家追溯厂商大包大揽,讲究的是一条龙服务、整套解决方案。而构件化的食品追溯系统运营模式区分为食品追溯系统构件生产商、食品追溯系统平台提供商、食品追溯系统个性化构件生产组装商以及食品追溯系统项目服务商等专业分工,做到分工明确,各司其职、各负其责,充分保护追溯链中各环节角色的利益。构件技术在食品追溯系统中的广泛应用,将使食品追溯系统做的更灵活、更经济、更易于推广。

4 结束语

纵观构件技术带来的诸多好处,亦能从中发现其难题所在就是领域构件的提取,还有就是领域构件的划分粒度问题,既要保证尽可能大的可重用性,又要保证尽可能小的功能可替换性。要解决这一问题,显然不能只靠软件方面的技术应用,而更重要的是要有食品安全追溯应用领域的丰富的领域知识,因此在领域工程的活动,不仅需要领域分析人员,还要有领域专家,提供关于领域中系统的需求规约和实现的知识,帮助组织规范的、一致的领域字典,帮助选择样本系统作为领域工程的依据,复审领域模型, DSSA 等领域工程产品。

从系统的架构实现上来说,目前国内一般的追溯系统的都是普通 B/S 的结构,不仅继承了传统 C/S 结构的优点外,还具有优越的系统性能、卓越的安全性能、减轻系统的负担、易于维护和升级等优点。但显然这些优点都只是相对的,在需求多变的领域在软件开发上其优势还是不够的,因此本系统以构件技术开发的 B/S 系统就应运而生。随之而来的问题就是这些构件之间的集成问题,毕竟不管构件开发的如何完美,都必须能集成一个可用可靠的应用系统,并要满足构件能够装配互换。由于分布式的服务器部属,安全追溯系统的有关网络传输方面不仅涉及传统的网络安全,还涉及基于可信性计算的安全构件。这些都是本系统的重要特点,因此成为需要重点关注及解决的问题。

参考文献

- [1] 王风云,赵一民,张晓艳,等.我国食品质量安全追溯体系建设概况[A].计算机农业应用分会论文选[C].2008第10期
- [2] 杨芙清,梅宏.构件化软件设计与实现[M].北京:清华大学出版社,2008

- [3] 裴庆裕, 耿玉水, 王新刚. 基于 JavaBean 的构件抽取和实现[J]. 山东轻工业学院学报 2008, Vol.22, No.1
- [4] 黄玉坤. 软件复用技术及领域工程综述[J]. 计算机与现代化. 2007 年第 11 期
- [5] 李克勤, 陈兆良, 梅宏, 等. 领域工程概述[J]. 计算机科学, 1999, Vol.26, No.5

作者简介

李骏平, 男, (1984 生), 计算机科学与技术专业软件工程方向的硕士研究生, 研究方向: 软件工程, 软件复用等。

王燕兴, 男, (1949 生), 教授, 硕士生导师, 研究方向: 软件工程, 软件复用及软件自动化等。

基于FPGA的UART的设计与仿真

李 琦 董利民 吴武臣 侯立刚

(北京工业大学 电子信息与控制工程学院, 北京 100124)

摘 要: UART 作为 RS232 协议的控制接口得到了广泛的应用。本文介绍了一种基于 FPGA 实现 UART 电路的方法, 并对系统结构进行了模块化分解以适应自顶向下的设计方法。采用有限状态机对各个模块进行了设计, 所有功能的实现全部采用 Verilog HDL 进行描述, 并在 Modelsim 6.0 环境下进行了仿真, 结果表明了该设计的正确性和可靠性。

关键词: UART; 有限状态机; FPGA

Design and Simulation of UART Based on FPGA

LI Qi DONG Li-min WU Wuchen HOU Li-gang

Abstract: UART is used widely as the interface of RS 232. This paper introduces a method to design UART circuit based on FPGA, and the system structure is divided into modularization to fit the design method of Top—Down. The modules are designed by FSM (Finite State Machine). All functions are described by Verilog HDL. We stimulate the functions under Modelsim 6.0 environment, the result proves the validity and reliability of the design.

Keywords: UART; FSM; FPGA

引言

随着微机系统的广泛运用和微机网络的极大发展, 通用异步收发器 (UART) 在数据通信及控制系统中得到了广泛运用, 其作用主要用来控制符合 RS232-C 协议的计算机与串行设备间的通信。常见的 UART 器件有 8250、NS16450 等芯片, 但是此类接口芯片存在体积较大、接口复杂以及成本较高的缺点, 而在实际应用中往往只需要 UART 的几个主要功能, 使用专用芯片会造成资源浪费和成本提高, 并且目前日趋成熟的 SOC 技术则要求将整个设计的功能集成在单片或几块芯片当中。因此, 将 UART 的功能集成在 FPGA 芯片当中, 满足相关系统需要, 提高了可靠性、稳定性和灵活性。本文提出了一种基于 FPGA 的 UART 的设计方法, 使用 Verilog HDL 语言^[1]开发, 利用有限状态机来描述 UART 核心控制逻辑的方法, 将其核心功能集成, 从而使整个设计更加稳定、可靠^[2]。

1 UART功能简介

UART 采用通用的 RS232-C 串行接口标准^[3]，该协议具有使用广泛的优点，在所有计算机和串行外设当中都设有这种接口，并且实现简单。UART 的具体帧格式如图 1 所示，UART 控制器所传输的一帧串行数据包括 1 位起始位，8 位数据位，1 位奇偶校验位和 1 位停止位。传输时，低位在前，高位在后。接收端检测并确认起始位后，接收 8 位数据位。停止位接收完毕后，向 CPU 发出中断信号，同时将数据送到计算机的 8 位数据总线上。发送数据时，先由 CPU 设置波特率，然后将 8 位并行数据加上起始位，奇偶校验位和停止位发送给外设。停止位发送完毕后，向 CPU 发出中断信号。在数据发送和接收过程中，CPU 可以加载控制信号来读取 UART 的工作状态，以便进行实时处理。



图 1 UART 的帧格式

2 UART的功能实现

本设计中，UART 主要包括采样模块、波特率产生模块、接收模块和发送模块 4 个部分。具体功能如下。

波特率产生模块：产生和 RS-232 通信所采用的波特率同步的时钟，以此才能按照 RS-232 串行通信的时序要求进行数据接收或发送，为系统内部提供相应时钟。

采样模块：在进行数据接收时，对从串行数据输入端口管脚发出的数据进行采样，提高数据准确性。

接收模块：接收从采样模块中发送来的异步数据，并进行串/并转换，发出中断信号，将数据发送到 8 位数据总线上。

发送模块：对从CPU 送来的并行数据进行并/串转换，发出中断信号，将8位数据添加起始位和停止位，并将数据串行发送到外部端口上。

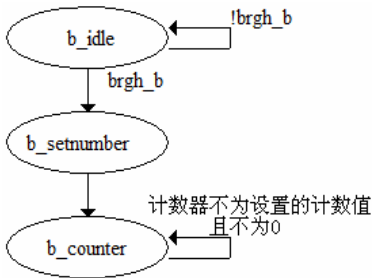


图 2 波特率产生模块状态转移图

1) 波特率产生模块

UART 的波特率由该模块决定，其实现原理是根据波特率的不同，对系统时钟进行分频。本模块通过编写状态机加以实现，状态转移图如图 2 所示。共包括三个状态，分别为：空闲状态(**b_idle**)，波特率设置状态(**b_setnumber**)和时钟翻转状态 (**b_counter**)。

在空闲状态 (**b_idle**) 下对所使用的计数器和对波特率设置寄存器进行清零设置。本设计中使用了 1 个 16 位的计数器 **counter** 作为时钟计数器。当 **brgh_b** 信号即波特

率产生模块的使能信号为高电平时,说明此时波特率产生模块已被触发,将转换到**b_setnumber**状态对波特率进行设置,若使能信号没有触发,状态机依旧保持在空闲态,等待下次触发。在波特率设置状态(**b_setnumber**)下,对波特率设置寄存器加以赋值。因为波特率产生模块的工作原理是对时钟进行分频,所以需要根据工作频率的不同,结合波特率,计算出时钟所需要的分频数 **N**。本模块所使用的波特率为 115600bit/s,具有波特率可调节的特点,设置其可以分别工作在 11M, 22M 和 40M 频率下,通过计算得出所对应的分频数分别为 6 分频, 12 分频和 22 分频。波特率设置状态完成后进入时钟计数状态(**b_counter**),通过时钟计数器控制时钟的反转,实现对时钟的分频产生内部符合波特率的时钟。

波特率产生模块仿真结果如图 3 所示。

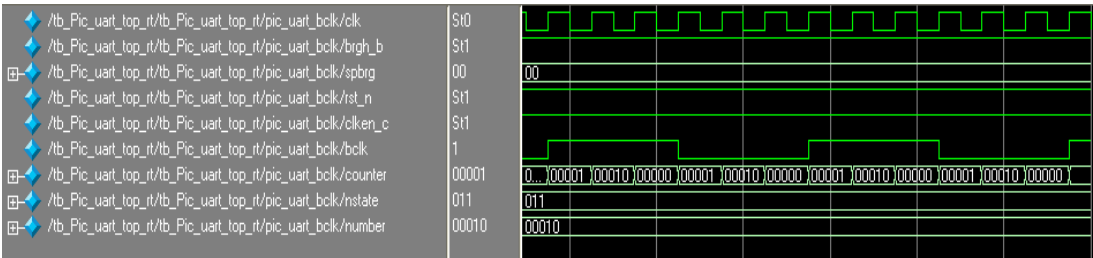


图 3 波特率产生模块仿真波形图

仿真结果说明:在设置 UART 工作在 11MHz 频率工作下时,clk 输入的时钟通过波特率产生模块,从 bclk 端口得到 6 分频的波特率时钟,说明所设计的波特产生模块能够正常工作,产生 UART 内部所需的波特率时钟。

2) 采样模块

为提高数据的准确性,本文所设计的 UART 在数据进行接收之前,需要对外部管脚输入进来的数据进行采样。本设计的采样方式为分别在内部时钟的 5,6,7 时钟沿的下降沿进行采样并作对比,对比标准为以 3 次采样,选择采样结果中 2 次相同的方式作为最终采样结果。

采样模块的设计的状态转换图如图 4 所示,其工作过程如下:在空闲状态(**sr_rec_idle**)下设置一个 16 位计数器 **scounter16** 对内部波特率时钟进行计数,赋初值为 4'b1110 并转换到采样状态(**sr_rec_sample**),在采样状态下对计数器进行减 1 操作,当 **scounter16**== 4'b1001 时,即在第 5 个时钟沿时,进行第一次采样,把此采样结果寄存在预先设置好的采样寄存器 **srx_pad_1** 中,计数器 **scounter16** 减 1,当计数器 **scounter16**== 4'b1000 时,第 6 个时钟沿到来,进行第二次采样,采样结果寄存在预先设置好的采样寄存器 **srx_pad_2** 中,计数器 **scounter16** 继续减 1,此时进行第三次采样,即在第 7 个时钟沿时执行相同的操作。把采样结果寄存在预先设置好的采样寄存器 **srx_pad_3** 中,采样状态完成后进入寄存器相加状态(**sr_rec_pulse**),此状态对 3 个采样寄存器进行相加,其相加结果作为采样结果的判断依据,在采样输出(**sr_rec_out**)状态中,根据上一状态中传送的结果,判断采样输出结果。本设计的基本判断依据为:当 3 个采样寄存器分别采的结果相加为 0,则说明 3 次采样都为 0,输出为 0 即 2'b00;若其中一个寄存器

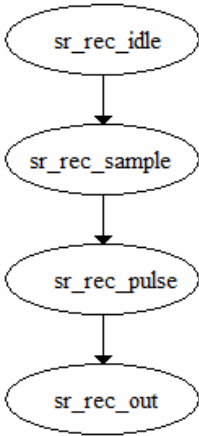


图 4 采样模块状态转换图

采样结果为 1 而其他寄存器为 0，则相加结果为 1 即 2'b01，依据“3 取 2”的原则采样输出结果为 0；若其中一个寄存器为 0 而其他寄存器为 1，则相加结果为 2 即 2'b10，采样结果输出为 1；最后，若所有寄存器采样结果都为 1，则相加结果为 3 即 2'b11，采样结果输出为 1。具体采样结果判断如表 1 所示。

表 1 采样结果判断表

采样寄存器 1	采样寄存器 2	采样寄存器 3	采样结果判断寄存器	采样结果
0	0	0	1'b00	0
0	0	1	1'b01	0
0	1	0	1'b01	0
0	1	1	1'b10	1
1	0	0	1'b10	0
1	0	1	1'b10	1
1	1	0	1'b10	1
1	1	1	1'b11	1

采样模块仿真结果如图 5 所示。

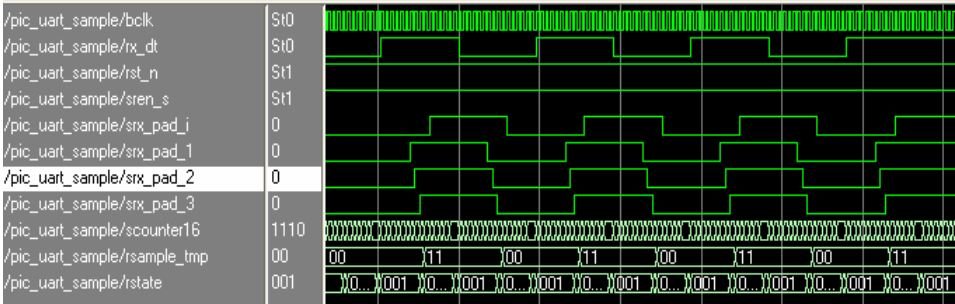


图 5 采样模块仿真波形图

仿真结果说明：通过仿真波形可以看出在 rx_dt 端口串行输入的数据，通过采样模块的 3 个采样寄存器 srx_pad_1, srx_pad_2, srx_pad_3 采样后，经过采样判断寄存器准确地从 srx_pad_i 端口进行输出。本设计的采样模块能实现准确的采样功能，提高了系统的准确性。

3) 接收模块

本接收模块设计为工作在异步模式下即串行数据帧和接收时钟是异步的，所以采集数据常用的方法有三倍速采样法，起始位中断捕捉、定时采样法。本设计的 UART 中的接收和发送模块均采用的是中间时刻采样法。通过使用一个内部计数器，对时钟进行 16 分频，在分频时钟的中点处对每一位数据进行传输，从而增加数据采集的准确性，提高稳定性。

本设计的模块由接收移位寄存器（RSR）和接收缓存寄存器（RCREG）两部分所组成。接收模块状态转换图如图 6 所示，在开始的空闲状态（sr_idle）中对内部计数器和中断信号寄存器分别赋值并侦测外部采样后数据的起始位，当采样的数据由高电平变为低电平时，说明外部端口已经开始传送起始位，状态转为接收起始位状态（sr_rec_start），在此状态下，先检验收到的数据是否为起始位，若数据由低变为高，则说明所接收的数据不是起始位，需回到空闲状态继续探测起始位，若数据没变化则转到准备接收状态（sr_rec_prepare）。准备接收状态中设置了接收数据位数和对 RSR 寄存器清零并转为数据接收状态（sr_rec_bit）。数据接

收状态接收了一位数据并将其送到 RSR 寄存器中,当 RSR 寄存器接收到一位数据并进行移位操作后,转到接收终止状态 (sr_end_bit)。在接收终止状态判断 8 位数据是否完全接收并移位到 RSR 寄存器中,若没有完全接收则返回到数据接收态继续执行接收操作,8 位数据全部传送到 RSR 寄存器后则转到接收停止位状态 (sr_rec_stop) 在此状态接收数据停止位,检测是否发生帧错误即数据是否接受完全。最后转到并行输出状态 (sr_push)。因为本设计的 RCERG 是一个两级寄存器,可以实现缓冲作用,进一步提高稳定性,加强数据接收的能力。因此需要先把一个 8 位数据传到第一个 RCREG 寄存器中,判断 RCREG 寄存器 1 中是否为空,若不为空则发出溢出错误信号 (OERR),反之,将 RSR 寄存器中的数据传送到 RCREG 寄存器 1 中并转到并行输出状态 1 (sr_push_1) 中,此状态依旧先检测 RCREG 寄存器 2 是否为空。若为空,则接受 RCREG 寄存器 1 传来的数据,并将 RCREG 寄存器 1 中的标志位置为空,返回到空闲状态 (sr_rec_idle) 等待接收下一个 8 位数据。

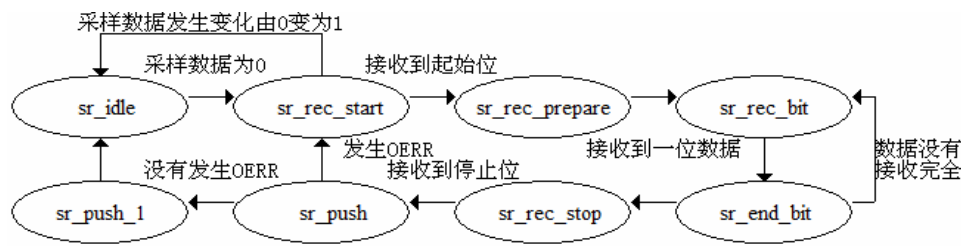


图 6 数据接收状态转换图

接收模块仿真结果如图 7 所示。

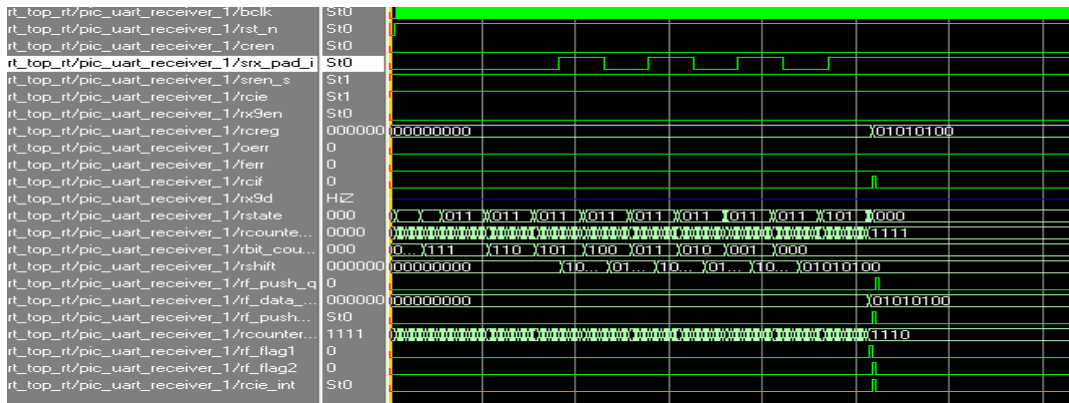


图 7 接收模块仿真波形图

仿真结果说明: 在仿真图中, 采样模块后的采样数据在 srx_pad_i 端口串行传送 01010100, 通过接收模块后, 并行发送到 RCREG 寄存器中, 发出中断信号, 说明仿真结果正确, 所设计的接收模块能够准确的接收数据。

4) 发送模块

本设计的发送模块由发送寄存器 (TXREG 寄存器) 和发送移位寄存器 (TSR 寄存器) 所构成。发送模块所设计的状态转换如图 8 所示, 共有 4 个状态。在空闲状态 (s_idle) 下通过负载信号 (txreg_load) 判断 TXREG 寄存器是否已装载数据总线上的 8 位数据, 若已装载, 向输出端口 (tx_ck) 发出高电平, 并转换到发送起始位状态 (s_send_start)。在发送起始位

状态 (s_send_start) 对内部计数器 counter 置零, 并对输出端口 (tx_ck) 持续发出起始位即逻辑“0”。进入串行发送状态 (s_send_byte), 通过移位操作将发送寄存器中的 8 位数据依次串行的发送到外部管脚寄存器 (tx_ck), 进入发送停止位状态 (s_send_stop)。通过对内部计数器的设置, 向外输出 1 位停止位, 即向输出端口输出逻辑高电平“1”, 发送中断 (txif) 并回到初始的空闲状态。

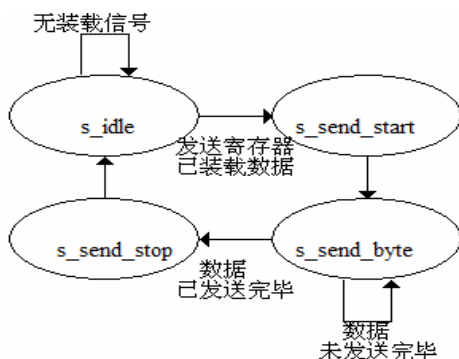


图 8 发送模块状态转换图

发送模块仿真结果如图 9 所示。

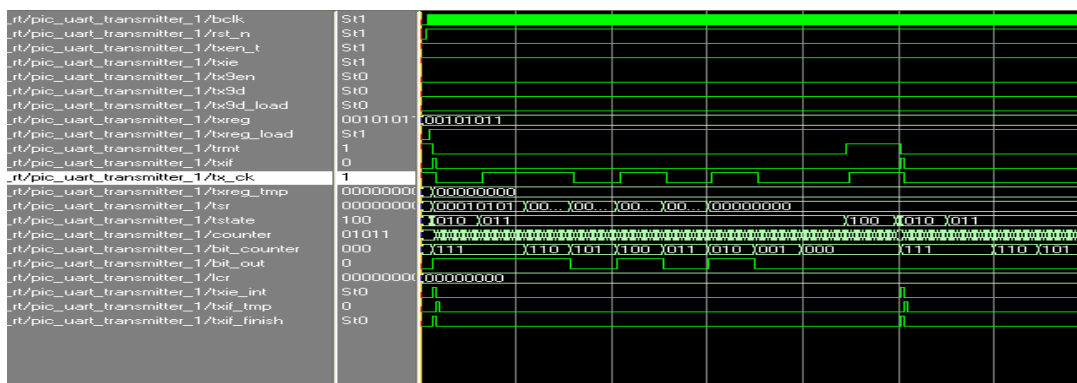


图 9 发送模块仿真波形图

仿真结果说明：在仿真图中，txreg 寄存器中从数据总线接收到的数据为 00101011，通过所设计的发送模块，在发送端口 tx_ck 得到串行输出数据 00101011，说明仿真结果正确，所设计的发送模块满足预期设计要求。

3 结论

使用 FPGA 实现 UART 模块的功能, 可以减小系统面积, 降低功耗, 同时使得设计更加紧凑和稳定。本文使用 Verilog HDL 语言在 xc2v1000fg256-4 FPGA 芯片上实现了 UART 模块的下载, 如图 10 所示。并通过 RS 232 接口连接到计算机, 使用串口调试助手软件在计算机上对 FPGA 进行数据的输入, 并且在显示器上直接观察到其输出的结果并通过验证^[4], 说明本设计的正确性和完整性, 各模块功能均达到预期的要求, 同时也表明本文的设计思想完全可行。



图 10 验证平台实物图

参考文献

- [1] Verilog HDL 程序设计与应用[M].王伟.北京:人民邮电出版社.2005,3
- [2] FPGA 系统设计与实践[M].黄智伟.北京:电子工业出版社.2005,8
- [3] 高级微型机计算机系统及接口技术[M].苏广川,沈瑛.北京:北京理工大学出版社.2006
- [4] Xilinx ISE Design Suite 10.x FPGA 开发指南[M].田耘,徐文波,胡彬等著.北京:人民邮电出版社.2008,11

作者简介

李琦, 1986 年 1 月 16 日, 硕士研究生, 研究方向为集成电路与系统集成。

运动图的方法计算合成双人运动

刘 宁

(北京工业大学计算机学院, 北京市多媒体与智能软件重点实验室, 北京 100124)

摘 要: 用运动图的方法实现双人抬木棒运动的计算合成。用光学运动捕捉仪捕捉运动数据, 然后对数据进行计算合成, 生成按照指定路径运动的新运动数据。对于运动数据合成的运动图方法进行了阐述, 并将运动图方法应用到了双人协同运动中。运动捕获数据包含了两个人身上共 72 个点在每一帧的三维坐标信息, 我对每一帧进行相似度计算, 然后生成运动图, 再对运动图进行搜索, 生成符合新路径的运动数据。

关键词: 运动图; 运动捕捉; 路径合成; 运动合成

Using Motion Graphs to Synthesize Two Men's Motion

LIU Ning

(Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: We use optical motion capture instrument to obtain the motion of two men lifting a stick. Then we calculate the data obtained, and generate new motion data which walk on new paths. Motion graphs as a common method to synthesis motions is introduced. We apply motion graphs to the two coordinated motion which are captured by the optical motion capture instrument. The motion data contains 72 points of the three-dimensional coordinate information per frame. We calculate the similarity of every two frames, and then generate the motion graphs. With searching the motion graphs we get new motion data which walks on new paths.

Keywords: motion graphs, motion capture, path synthesis, motion synthesis

1 概述

最近, 人体运动捕捉数据正广泛的应用于动画产业和游戏产业。和传统的手工制作的人体动画相比, 捕捉数据具有更高的真实感。人体运动捕捉数据比较完美地记载了人体运动的各种细节。一般来说, 运动捕捉数据是比较难修改的, 一旦修改不慎, 就会使数据的真实性

资助项目: NSFC-广东联合基金 (U0935004), 北京市教委科技创新平台项目。

打折扣。如果现有的数据不能满足要求,只能是重新去采集满足要求的数据。而采集运动数据是一项昂贵且费时费力的工作。对运动捕捉数据进行计算合成,同时保留数据的真实性是比较困难的。

中科院陆汝钤院士首次提出了全过程计算机辅助动画自动生成技术^[1,2],该项技术是人工智能与图形学技术相结合的产物,近年来人工智能技术在很多领域都得到了应用,如人机对战,专家系统等。中国古典建筑是中华文化的瑰宝,现存的古典建筑所体现出来文化内涵是五千年历史的积淀。随着计算机产业的发展,计算机技术已经应用到我们生活中的方方面面,但中国古典建筑信息化方面所做的研究却少之又少。本项目组在计算机辅助动画自动生成技术研究的基础上,将这一技术应用到古典建筑大木作结构的研究中来,做出了计算机技术应用到中国古典建筑领域的初步探索,对于我国物质文化遗产的保护具有重要意义。在古建场景中,双人抬木棒的运动是比较常见的,本论文的方法用于生成适合新的古建场景的双人运动数据。

2 运动图的一般方法

Schodl 和他的同事提出了一种方法^[3],用于将一段短视频片段合成为一段长视频片段,受到此方法的启发,运动图(motion graphs)的方法在2002年被很多研究小组同时提出,并在以后的时间里,不断得到完善和改进。无论是基于交互控制的人体运动数据合成^[4,5]还是基于草图的人体运动数据合成^[6,7],运动图都是一种很有实用性的方法。

在一组运动捕获数据片段的基础上就可以生成一个运动图。运动数据中的每一帧都可以看做是运动图中的结点,如果从一个结点可以平滑的过渡到另一个结点,运动图中就对应有一条连接这两个结点的边。平滑过渡指的是两帧数据连在一起组成一个流畅的运动。

首先,定义一个矩阵来找到可以平滑过渡的结点。一般的方法是比较两个结点A和B(也就是两帧数据)之间的相似度,如果相似度低于用户定义的一个阈值,A结点就可以连接到B+1结点,B结点可以连接到A+1结点(A+1结点和B+1结点分别是A结点和B结点在各自运动数据片段中的后继结点)。我们使用了Kovar et al.^[8]的点云矩阵来计算两帧之间的相似度,在判断相似度方面,这是一种比较直观和简便的方法。计算每两帧之间的相似度,在用户规定好阈值后,只取相似矩阵中低于阈值的局部最小值作为可以连接的结点。根据相似矩阵的计算结果,生成一个图,结点是运动捕获数据的一帧,边可以是已经采集的运动数据,也可以是相似矩阵计算得到的连接两个相似结点的边。然后计算生成的图的最大连通子图(the largest strongly connected component),把图中不是连接该最大连通子图中两个结点的边全部删除。当然,如果生成了两个或几个包含结点数相当的最大连通子图,也可以都保留他们。运动图生成以后有很多具体的应用,比如说用人体走路的动作进行计算得到一个运动图,给定一条路径,在运动图中进行搜索,如果运动图的连接性足够好,就可以得到一段新运动数据,使这段数据沿着事先给定的路径行走。新生成的数据基本保持原采集动作的真实性。还可以根据运动图来进行交互控制,实时的根据用户的输入来决定搜索图的哪一部分,从而能进行实时控制。

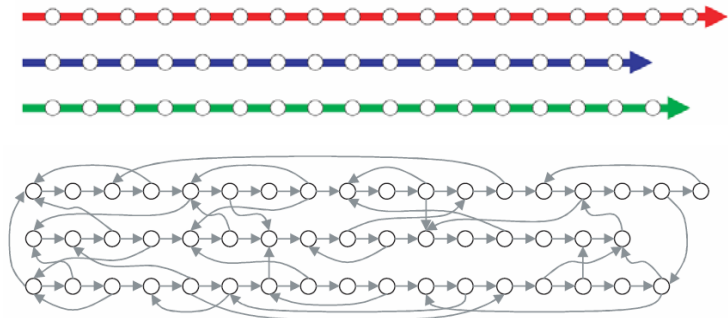


图 1 上半部分为原始的运动数据，下半部分为经计算得到的运动图

3 双人运动图

本文用的实验数据为两人抬木棍的运动，这种动作在古代工地也是比较常见的。因为所使用的光学运动捕捉仪的场地比较小，长宽大概分别为四米和三米，所以能采集到的两人运动数据都是一些在较短路径上运动的数据。采集设备使用.trc 格式的文件保存数据，采集的帧率为每秒钟 50 帧。设备有 22 个摄像头，两个模特每个人身上都有 36 个贴有漫反射材质的小圆球。设备采集到的每帧的数据就是这 72 个小原点的三维坐标。trc 文件内容还包括每帧的时间戳，以及包含的总帧数等。两人抬木棍运动我总共采集了 82 个片段，总共 2 万余帧。

trc 文件是我们所用的光学采集设备的专用格式，这种格式可以通过 autodesk 公司的软件 motionbuilder 转化为 bvh 格式文件。bvh 格式是 maya, 3ds max 等动画软件制作人体骨骼动画时常用到的文件格式。这样我们计算的得到的数据经过处理后可以用于动画的制作。

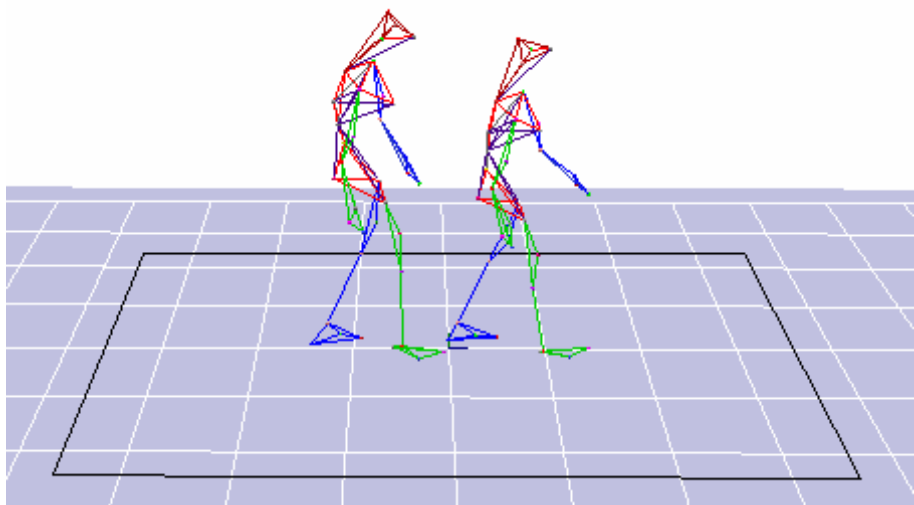


图 2 双人抬木棍运动的一帧，木棍没有显示，黑色矩形框为实验设备的有效采集范围

我编写的系统的输入为 trc 文件，输出为经过运动图处理运算合成后的 trc 文件。计算所用的坐标系满足右手准则，x 轴 z 轴构成水平面，y 轴方向为 xz 的右手准则方向。

1) 计算相似度

对任意帧 A 和帧 B 计算它们的 72 个对应点之间的距离来判断这两帧是否相似，是否可以给这两帧在运动图中创建一条边。因为这两帧在世界坐标系中的相对位置是不一样的，朝向也不相同，所以要计算他们的对应点的距离，首先要将其中一帧进行旋转和平移，使这一帧与另一帧的位置重合。具体方法是：对 A 中第一个人的 36 个点的 x 坐标和 z 坐标取平均值，设为 a1，对 A 中第二个人的 36 个点的 x 坐标和 z 坐标取平均值，设为 a2，对 B 中两个人也分别去平均值，为 b1 和 b2。设由 a1 指向 a2 的二维向量为 m1，由 b1 指向 b2 的二维向量为 m2，求得 m2 与 m1 的夹角，使用该夹角构造旋转矩阵 M。对 A 帧的 72 个点的 xz 坐标取平均值，对 B 帧的 72 个点的 xz 坐标去平均值，这两个平均值之差设为 x，z。这样 B 帧乘以矩阵 M 然后将 72 个点的 x 坐标和 z 坐标分别加上 x，z，得到的帧就是和 A 帧大体重合的，这时利用三维坐标距离公式计算对应点在世界坐标的距离，就能较好的判断这两帧的相似度。对每两帧进行计算，得出帧 A 到帧 B 所需要的旋转角度，x 坐标的平移值和 z 坐标的平移值，和这两帧之间的距离。

2) 生成运动图

① 选择合适的连接点

对每两帧进行相似度计算，得到一个距离矩阵，这个距离矩阵的局部最小值作为过渡结点的候选点。一个距离矩阵的局部最小值并不一定是一个高质量的连接点，选定一个阈值，当局部最小距离小于这个阈值时，才作为生成的运动图中的连接点。这个阈值过高，连接点的过渡性就不好，如果太低，就会使连接点的数量过少。

② 剔除图中的无用结点

经过这一步后生成的图中有很多没有用的结点，比如有些结点的时间戳的太近，例如计算得到 A 帧和 B 帧相似，然而他们的时间戳只相差不到一秒，这样的结点是应该作为相似结点的。还有一些结点是死结点，当遍历图的这个结点时，将没有后继结点或者在局部的很少结点进行死循环，这样的结点也应该从运动图中删除。我使用 Tarjan 算法来生成强连通子图（SCC）。把强连通图以外的结点删除。如果图中一条边的两个结点中的一个或两个不在生成的强连通图中，也把这条边删除。经过这样的处理后，运动图中剩下的任何一个结点与其他结点都是连通的。

这样便得到了最终的运动图。图中的边有两种类型：第一种，边的两个结点在原始运动数据中就是连在一起的，它们之间有可能有一帧或很多帧；第二种，边的两个结点是我们经过相似计算得到的相似结点。对于第二种边，它要包含一些信息，这些信息就是我们在计算它的两个结点的相似度时得到的旋转信息，x 轴平移量和 z 轴平移量。

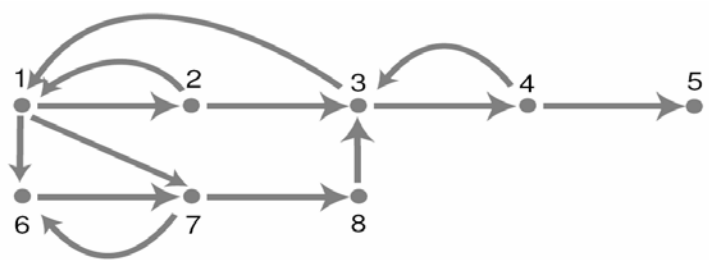


图 3^[8] 一个未经处理的运动图。最大强连通分量为[1,2,3,6,7,8]。4 是一个冗余结点，5 是死结点

3) 运动图的应用

从运动图的某个结点开始,沿着边进行深度优先的搜索,经过的这些边经过计算就可以合成一段新的运动数据。运动图中的一个结点,对应原始 **trc** 文件中数据的一帧,包含两个人身上 72 个结点的三维坐标信息。假设已经在运动图中搜索到一条路径,则根据此路径生成新运动数据的算法描述如下:

```
myStack; //自定义的栈,用来保存图的行走中碰到的第二种边中的数据信息
//栈的每个元素包含
//angle (旋转角)
//x (x轴平移量)
//z (z轴平移量)
while (没有到达结束结点)
{
    if(边是第二种类型)
        {该边包含的信息入栈}
    if (边是第一种类型)
        {
            指针p指向栈顶
            while (指针p没有指到栈底)
            {
                边的每一帧乘以angle构成的旋转矩阵
                边的每一帧加上平移量x
                边的每一帧加上平移量z
                在trc文本中打印一行数据
                指针p指向栈的下一层
            }
            结点指针指向下一个结点
        }
}
```

如果生成的运动数据需要有整体的旋转和平移,那么先将这个信息入栈,接着进行上面的算法即可。

经过上面的算法,就可以将一条图的搜索路径转化为一段新生成的运动数据,用 **trc** 文件保存。

4) 运动路径合成

从理论上讲,使用的运动捕获数据越多,经过计算生成的运动图的结点就越多,对图的路径进行搜索,能够合成的更多的新的运动数据。我们采用的双人抬木棍的运动生成运动图,如果已知双人运动的路线,那么运动图中肯定存在一段运动路线最接近该路线的图路径。合成新路线运动,实际就是一个图的搜索问题。我使用了一个局部搜索的办法来实现运动路径的拟合。从图的某一个起点开始搜索(这个起点要经过旋转和平移,使两个人的运动方向与路线的起点切向量相同),选择该结点的子结点中最靠近需要拟合的路线的点作为结果结点,以此类推,再选择下一个最近的点作为结果结点。当离路线终点的距离小于一个阈值时,搜索结束。然后运用第 3 节的算法,由图的结点序列生成新的运动数据。经过实验,本方法实现了较好的合成效果。图 4 展示的是新合成的双人沿圆形路径的运动。

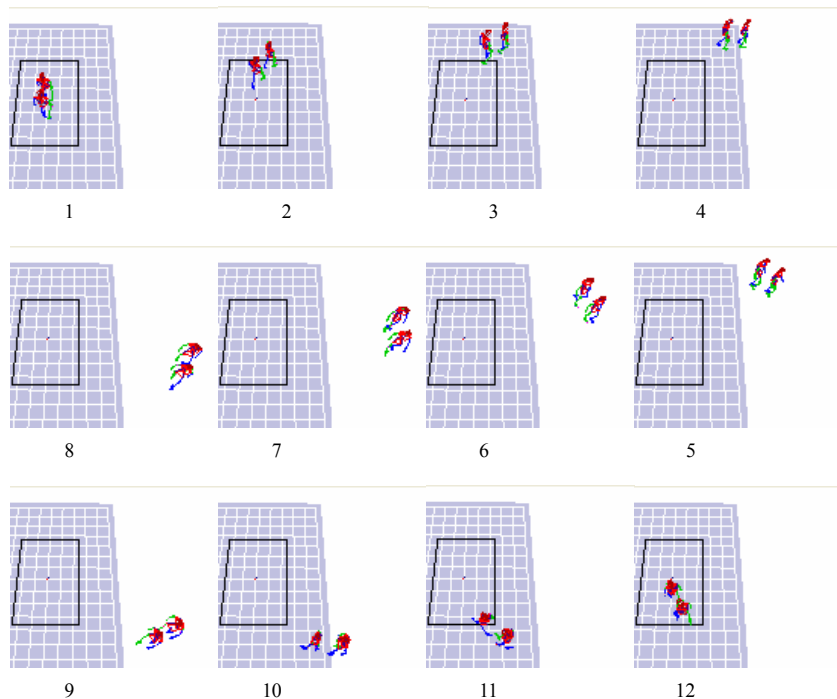


图 4 指定圆形路径后和合成的运动数据

5) 方法的不足和下一步的改进

本文的双人运动图方法参考了 KOVAR 的单人运动图方法^[8], 将其运用到双人运动。在具体的计算方法上, 有很大不同。本文的方法对数据质量的依赖程度还非常高, 两个人的运动不像一个人运动那样容易寻找相似点, 因为两个人一起运动, 胳膊和腿的数量比一个人加了一倍, 而且计算相似度时还要考虑两个人的相对位置。现在对路径的拟合程度现在只能达到比较粗的粒度, 下一步研究是找到更好的判断相似点的方法, 以及运动的编辑方法, 在符合运动约束下进行反向运动学计算, 使动作衔接更加流畅, 使路径的拟合能够更加精细。

参考文献

- [1] Lu Ruqian, Zhang Songmao. Automatic Generation of Computer Animation[J]. Springer-Verlag, 2002.33-35.
- [2] 陆汝铃, 张松懋. 从故事到动画片-全过程计算机辅助动画片自动生成[J]. 自动化学报, 2002,28(15).321-348.
- [3] SCHODL A., SZELISKI R., SALESIN D. H., ESSA I.. Video textures[C]. In Proceedings of ACM SIGGRAPH 2000(July 2000), Computer Graphics Proceedings, Annual Conference Series. 489-498.
- [4] LEE J., CHAI J., REITSMA P. S. A., HODGINS J. K., POLLARD N. S.. Interactive control of avatars animated with human motion data[J]. ACM Transactions On Graphics, 2002, 21(3). 491-500.
- [5] KWON T., SHIN S. Y.. Motion modeling for on-line locomotion synthesis[C]. In ACM SIGGRAPH/Eurographics Symp. On Comp. Animation. July 2005. 29-38
- [6] ARIKAN O., FORSYTH D. A.. Interactive motion generation from examples[J]. ACM Transactions On Graphics, 2002, 21(3). 483-490.

- [7] SAFONOVA A., HODGINS J. K.. Construction and optimal search of interpolated motion graphs[J]. ACM Transactions On Graphics, 2007, 106-113.
- [8] KOVAR L., GLEICHER M., PIGHIN F.: Motion graphs[J]. ACM Transactions On Graphics, 2002, 21(3), 473-482.

作者简介

刘宁，男，1982 年 3 月 16 日出生，研究生，山东省莱芜市人，研究方向为计算机图形学。

全过程计算机辅助古代塔类建筑动画自动生成

刘射彪 梁天柱

(北京工业大学计算机学院, 北京市多媒体与智能软件重点实验室, 北京 100124)

摘要:《基于语义理解的古建动画辅助生成系统》中的《塔类建筑三维动画生成平台》依据使用者输入的对古代塔类建筑的描述, 生成符合使用者描述的古塔类建筑的 3D 模型。整个借助计算机的辅助完成, 与传统的手工动画制作方法相比, 在效率方面具有明显的优势。并且对将古代建筑知识的普及, 有其重大现实意义。

关键字: 动画自动生成; 三维动画; 塔类建筑; 人工智能

Full Life-Cycle Computer Aided Pagoda Animation Generation

LIU She-biao LIANG Tian-zhu

Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology, Beijing 100124, China

Abstract: Depending on the inputs of uses, the *3D Animation Generation Platform for Pagoda in Semantic-Based Computer Aided Ancient Architecture Animation Generating System* could generate 3D-modules of the pagodas accord with the description of users. The whole life-cycle of the generation processing is aided by computer. Compare with the handmade way of animation manufacturing, the system has an obvious advantage in efficiency. Nevertheless, for the popularization of knowledge of ancient architecture, there are great practical significance.

Keywords: Automatically Animation Generation, 3D Animation, Pagoda, Artificial Intelligence

1 引言

伴随着计算机科学技术的不断发展, 计算机技术已经逐渐融入人类生活的方方面面。在动画制作过程中, 也不乏计算机技术的身影。《功夫熊猫》、《机器人总动员》等动画影片, 代表了近些年来计算机动画 (Computer Animation)^[1,2]的最高水平。当今动画技术已经应用于动画故事片、影视、广告、电脑游戏以及虚拟现实模拟等众多领域, 尽管计算机辅助动画已经取得了长足的发展, 但是很多大量重复耗时的工作仍然是由动画制作人员亲自完成, 生产效

基金项目: 国家科技支撑计划中的古代建筑虚拟修复及 Web 表现技术研究课题。

率也受到了很大的影响。在 20 世纪 90 年代中科院陆汝钫院士首次提出了全过程计算机辅助动画自动生成技术^[3,4]，该项技术是人工智能与图形学技术相结合的产物，是人工智能技术的一个新的试验场，是人工智能技术在计算机图形图像等方面的新应用，从动画产业的发展的长远来看，动画自动生成技术能够大大提高动画制作人员的工作效率，降低动画制作成本。

古代建筑是中国古老文明的象征，是中国文化中的一颗璀璨的明珠，但在中国历史上见于文字记载的建筑中，能够保留下来的比例极低，今天依然能够使用的更是屈指可数。随着大量西方风格的摩天大楼、住宅社区遍地开花，如何保护古老建筑的问题日益引起关注。

目前，我国在古建筑知识保护领域已经逐步采用了图形学的先进技术，将大量宝贵的古建筑知识数据化，电子化，对古建筑知识普及推广起到很大的作用。三维模型及其动画展示是其中最常采用的方法。在这样的大环境下，北京工业大学和中国科学院数学与系统科学研究所联合为湖南省博物馆开发了《基于语义理解的古建动画辅助生成系统》。

在此之前，中国建筑科学研究院建筑工程软件研究所开发了一款中文三维图形平台——《中国古典建筑设计软件》（PKPM-GUCAD）。该软件通过对中国古典建筑的各种营造法式和做法则例的分析，将其中规律编入计算机程序，自动生成各类古建筑模型。文中介绍的《基于语义理解的古建动画辅助生成系统》（下简称《古建动画系统》）与之相较，有其特有的优势，《中国古典建筑设计软件》面向的受众是有具有相关古建专业知识的古建工作者，其操作上有一定的复杂性和较强的专业性，对使用者的专业知识的积累有较高的要求。而《古建动画系统》面向的使用群体则更为广泛，既面向古建专业工作人员以及古建领域专家，也面向广大的古建爱好者，甚至是普通的不具备古建知识的使用者。《古建动画系统》的操作界面如图 1 和图 2 所示。



图 1 基于语义的古建动画辅助生成系统网络版界面



图 2 基于语义的古建动画辅助生成系统线下版界面

2 塔类建筑三维动画生成平台

1) 简介

《塔类建筑三维动画生成平台》（下简称《塔类建筑平台》）是《古建动画系统》的重要组成部分。在《塔类建筑平台》中，集成了我国丰富的塔类建筑形制中具有代表性的六种常见的、各具特色的塔类建筑，其中包括：楼阁式塔、密檐式塔、花塔、覆钵式塔（也称喇嘛塔）亭阁式塔以及金刚宝座塔。

《塔类建筑平台》的系统设计方案保证了对于建筑构件规则以及模型库的可扩充性。使用者可以根据自己的需要，添加新类型的塔的部件模型及其搭建规则以丰富系统。



图 3 塔类建筑三维动画生成平台所涵盖的六种塔类建筑的实物图例（图片来自互联网）

2) 系统流程

系统流程描述：

a) 通过用户的输入，得到对塔类建筑的分类及其相关属性的首先自然语言的描述。

b) 通过核算模块生成建筑部件定量信息文件 ToADL.xml（CAL 的一部分，描述模型部分的），获得所有运动主体的信息。由这些参数信息，尤其是搭建顺序信息，借助建筑动作库，

通过动作规划生成粗粒度定性描述 ADL。

c) 根据 ToADL.xml 中的部件信息，获得同类建筑部件分组情况，将粗粒度动作描述细化，通过动作二次规划生成细粒度动作定性描述 ADL。

d) 基于模型库和 ToADL.xml 文件，通过动作计算模块将动作定性描述信息量化，生成动作定量描述语言（CAL）。

e) 由定量描述语言 CAL，通过动画文件生成模块生成包括所有运动信息和模型信息在内的可演示的动画文件。

系统流程图如图 4 所示。

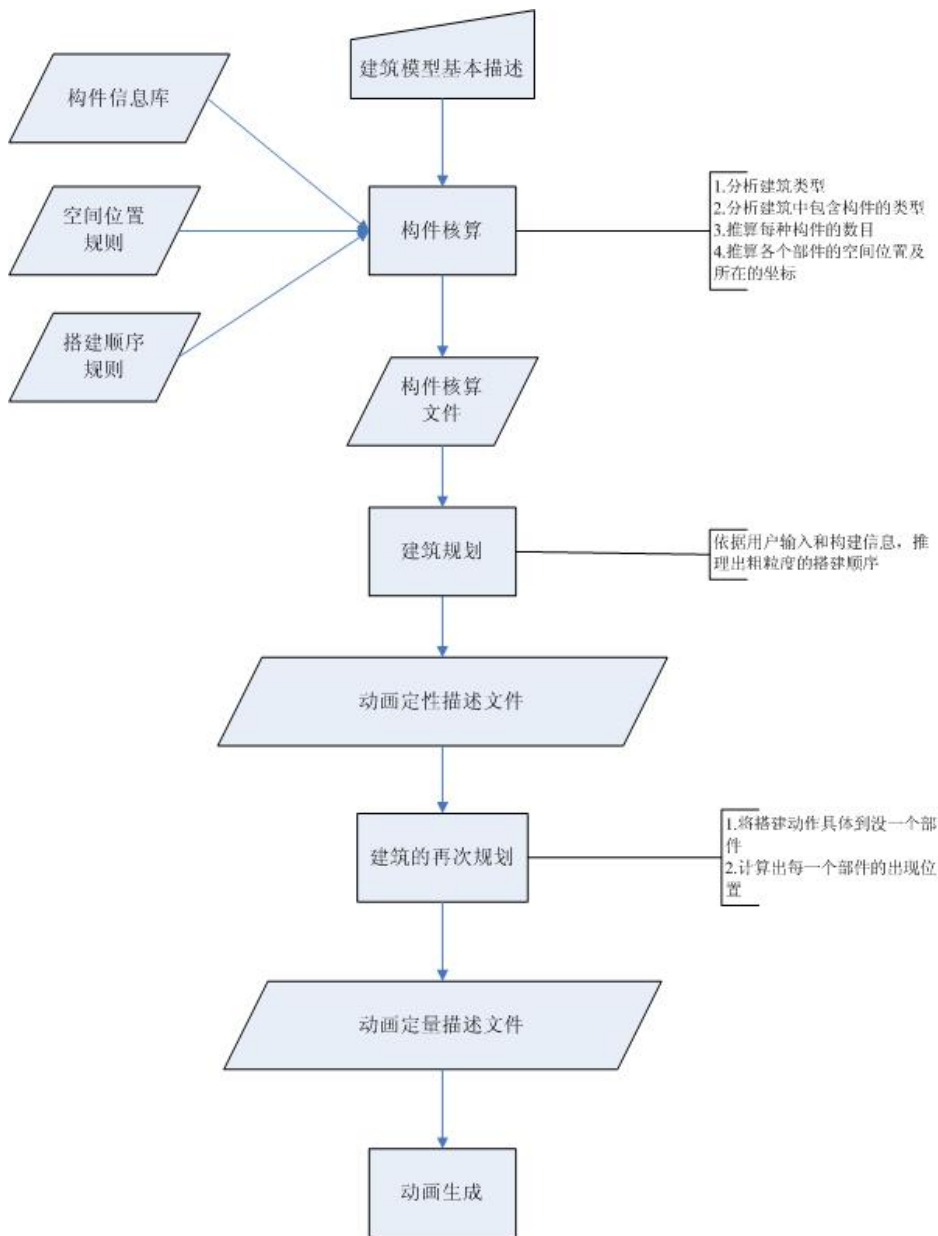


图 4 塔类建筑三维动画生成平台的流程图

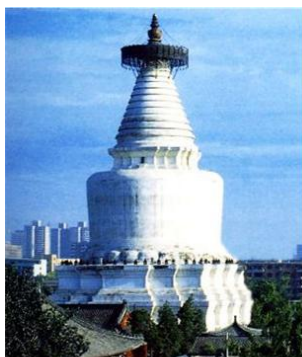
3) 系统实现

本节以塔类建筑中的覆钵式塔（喇嘛塔）为例，介绍塔类建筑三维动画生成平台的具体功能实现。

① 知识介绍

到了元代，窣堵波再次传入中国，演化为覆钵式塔，覆钵式塔的特征非常明显，它的塔身部分是一个半圆形的覆钵，在其上安置高大的塔刹。覆钵之下，建一个巨大的须弥座承托。半圆形的覆钵还基本上保存了坟冢的形式。因其为喇嘛教常用建塔形式，故也称为喇嘛塔、藏式塔^[5,6]。

我国现存的覆钵式塔有，北京妙应寺白塔，北京北海琼岛白塔，五台山塔院寺白塔，以及扬州瘦西湖莲性寺白塔等。实物图如图5所示。



北京妙应寺白塔



北京北海琼岛白塔



山西五台山塔院寺白塔

图5 我国现存的几座著名的覆钵式塔（图片来自互联网）

② 参数限制

古代建筑在搭建过程中需要遵循相应的营造法式与建造规则，尽管塔类建筑在样式上较为松散，风格迥异，但在其搭建思想中在变化中仍然存在一些不变的内容。而在塔类三维动画生成平台中通过查阅资料和对现存各种塔类建筑实例的分析与归纳，总结出一套相对完善的塔类建筑参数设定规则。

通过对建筑文献资料以及各种实物图的分析 and 总结，针对各种形制不同的覆钵式塔剥离出四个具有代表性的可变内容，分别为：金刚圈的类型、是否有双耳装饰、伞盖的类型以及宝刹的类型。覆钵式塔的结构示意图如图6所示。

针对这四个可变内容选择的输入限制如下。

a) 金刚圈的类型：可选的内容有“善释塔金刚圈”、“降佛塔金刚圈”、“涅槃塔金刚圈”、“和解塔金刚圈”、“胜利塔金刚圈”、“法轮塔金刚圈”、“神变塔金刚圈”、“菩提塔金刚圈”等八种不同类型的金刚圈。

b) 是否有双耳装饰：可选择的内容有“有双耳装饰”或者“无双耳装饰”。

c) 伞盖的类型：可选择的内容有“流苏华盖”和“天地盖”。

d) 宝刹类型：可选择的内容有“日月刹”、“金属高刹”和“宝珠刹”。

③ 部件空间位置、旋转及尺寸计算规则

a. 空间位置计算：

在知识总结的过程中，将塔类建筑部件在位置计算中的情况抽象为几组具体的公式，可

以用于计算各种不同类型的部件。覆钵式塔的整体结构是自底向上，依次叠加的（如图 7 所示），计算涉及计算公式如下：

x 轴方向的坐标： $positionX = 0$ ；（x 轴方向的偏移量为 0）

y 轴方向的坐标： $positionY = H + \frac{height}{2}$ ；

z 轴方向的坐标： $positionZ = 0$ （z 轴方向的偏移量为 0）

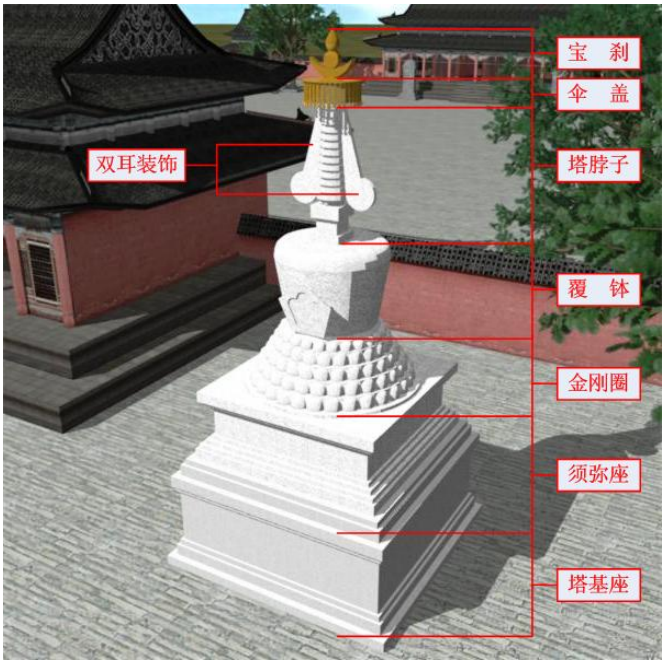


图 6 覆钵式塔结构示意图

其中 $positionX$ 、 $positionY$ 和 $positionZ$ 分别表示部件位置的 X 轴、Y 轴和 Z 轴的坐标， $height$ 表示当前部件的高度（Y 轴方向）， H 表示当前部件的高度基准（具体，实际数据）。

下面以计算覆钵式塔中的部件“覆钵”为例，介绍计算公式的使用法。“覆钵”的前驱部件为“金刚圈”（图 7 中金刚圈的类型为“善释塔金刚圈”），金刚圈的上顶面高（y 轴方向坐标）为 154.0，则“覆钵”的高度基准 $H=154.0$ ，覆钵自身的高度 $height=56.0$ ，则“覆钵”的空间位置坐标中的 x 轴和 z 轴方向坐标均为 0.0，y 轴方向坐标为 $154.0+56.0/2=182.0$ ，即覆钵的位置为（0.0，182.0，0.0）。

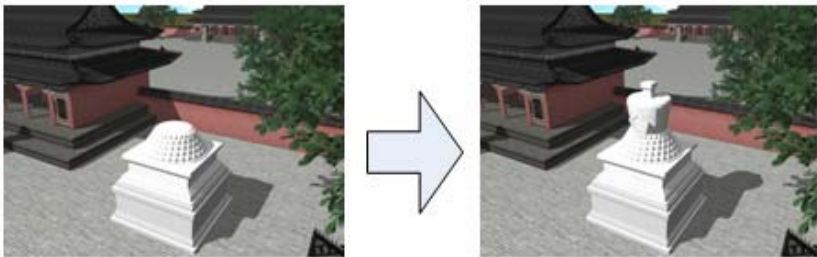


图 7 覆钵式塔生成示意图

b. 空间旋转角度

在知识总结的过程中,将塔类建筑部件在空间旋转角度计算中的情况抽象为几组具体的公式,可以用于计算各种不同类型的部件。覆钵式塔的整体结构是自底向上,依次叠加的(如图7所示),用到的是针对无需旋转的部件使用的计算公式,计算公式表述如下:

绕 x 轴旋转的角度: $\text{rotateX} = 0$;

绕 y 轴旋转的角度: $\text{rotateY} = 0$;

绕 z 轴旋转的角度: $\text{rotateZ} = 0$

其中 rotateX 、 rotateY 和 rotateZ 分别表示部件绕 X 轴、Y 轴和 Z 轴的旋转的角度。

b. 空间尺寸计算

一些塔类建筑在建造上存在随着层数的递增而逐层按比例缩放的问题,这个问题主要针对的是多层结构的塔,如楼阁式塔和密檐式塔。

设 scale 为缩放比例,则位于第 n 层的构件其尺寸的计算公式应为:

$\text{sizeX} = \text{width} * \text{scale}^{n-1}$;

$\text{sizeY} = \text{height} * \text{scale}^{n-1}$;

$\text{sizeZ} = \text{depth} * \text{scale}^{n-1}$

其中 sizeX 、 sizeY 和 sizeZ 分别表示部件在 X 轴、Y 轴和 Z 轴方向的尺寸,即部件的宽度、高度和厚度。

空间尺寸计算主要用于楼阁式塔、密檐式塔及花塔这种多层次结构的塔,在覆钵式塔的结构中并不涉及部件的缩放问题,因此在覆钵式塔的部件尺寸计算中没有用到上述的几个公式。

④ 顺序规则

(1) 规则描述

首先建立虚拟的塔类建筑结构,以楼阁式塔为例,楼阁式塔从结构上看,首先有一个总体的建筑结构——塔,依据楼阁式塔的建筑特点,这个总体的建筑结构可以被切分为多个层,具体的层结构可以划分为若干构件,如:塔墙、塔身柱、塔檐等。在楼阁式塔中的层次间往往都是重复的。可以将其顺序规则划分为三个不同的类型,楼阁式塔基层部件,楼阁式塔上层部件(包括第二层到倒数第二层之间的各层)以及楼阁式塔顶层部件。针对上述三类部件建立不同的规则。

下图以覆钵式塔为例,介绍如何利用“树形结构”来描述塔的结构,描述如下。

塔类建筑的搭建过程并不像法式建筑那样有固定的依据可循,其搭建顺序更是无法可依,因此为了便于塔类建筑搭建过程的展示,其搭建顺序采用的是默认自底向上,由低到高,分模块搭建,这样的生成过程便于观看者了解塔类建筑的结构,便于其了解各种类型的塔的特有属性。在顺序结构的描述方面,通过对树的广度优先遍历来确定部件的搭建顺序以及塔的层次结构。

(2) 适用性

该描述方法的适用性比较广泛,其描述能力涵盖了各种类型的塔类建筑,不仅可以用来描述多层次的塔类建筑,也可以用来描述如亭阁式塔这类单层结构的塔,单层塔是多层塔的一个特例。

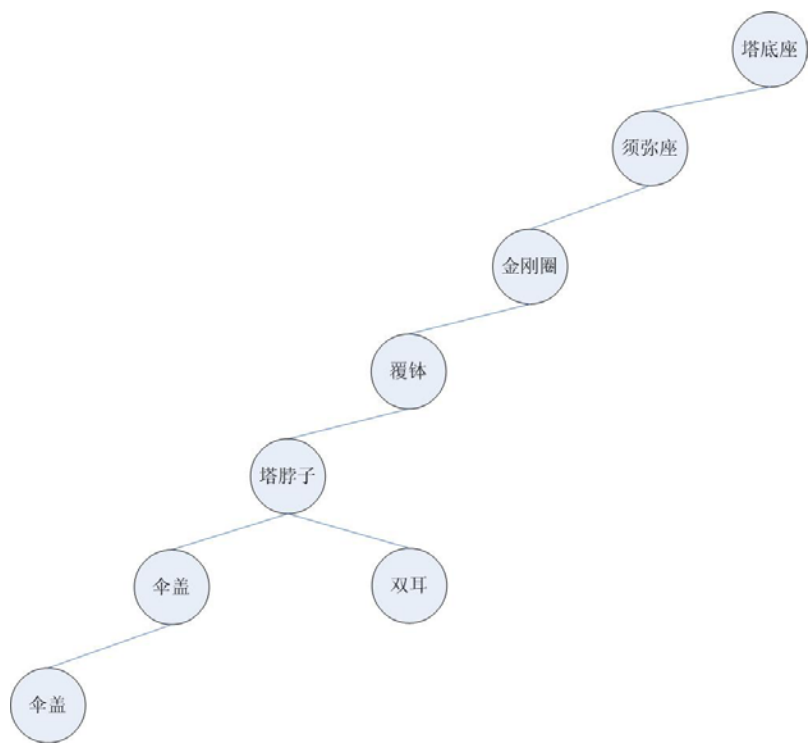


图 8 覆钵式塔的树形结构实例

3 生成结果展示

1) 系统实现

基于使用者输入不同的参数,《塔类平台》能够生成不同形制的塔类建筑共计 144 种,统计明细表 1 所示。

表 1 《塔类平台》生成结果统计表

塔类形名	生成的种类数目
楼阁式塔	10 种
密檐式塔	18 种
花塔	6 种
覆钵式塔	96 种
亭阁式塔	12 种
金刚宝座塔	2 种
共计	144 种

2) 生成过程展示

① MAYA 动画版本

生成的模型为一个描述参数为一个具有善释塔金刚圈,流苏华盖,日月宝刹,有双耳装饰的覆钵式塔。

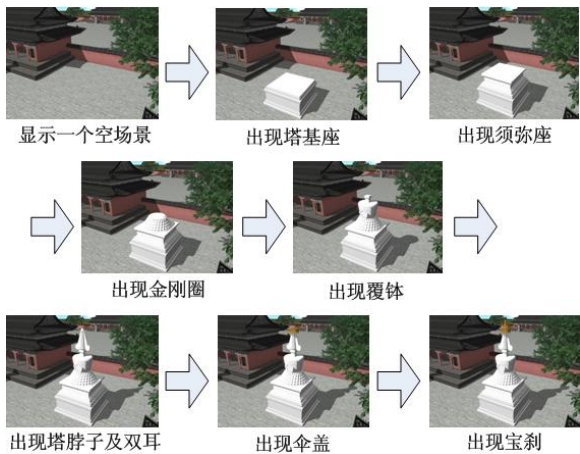


图 9 建筑生成动画展示 (MAYA 版本)

② VRML 版本

生成的模型为一个 8 角的，层数为 5 的楼阁式塔。

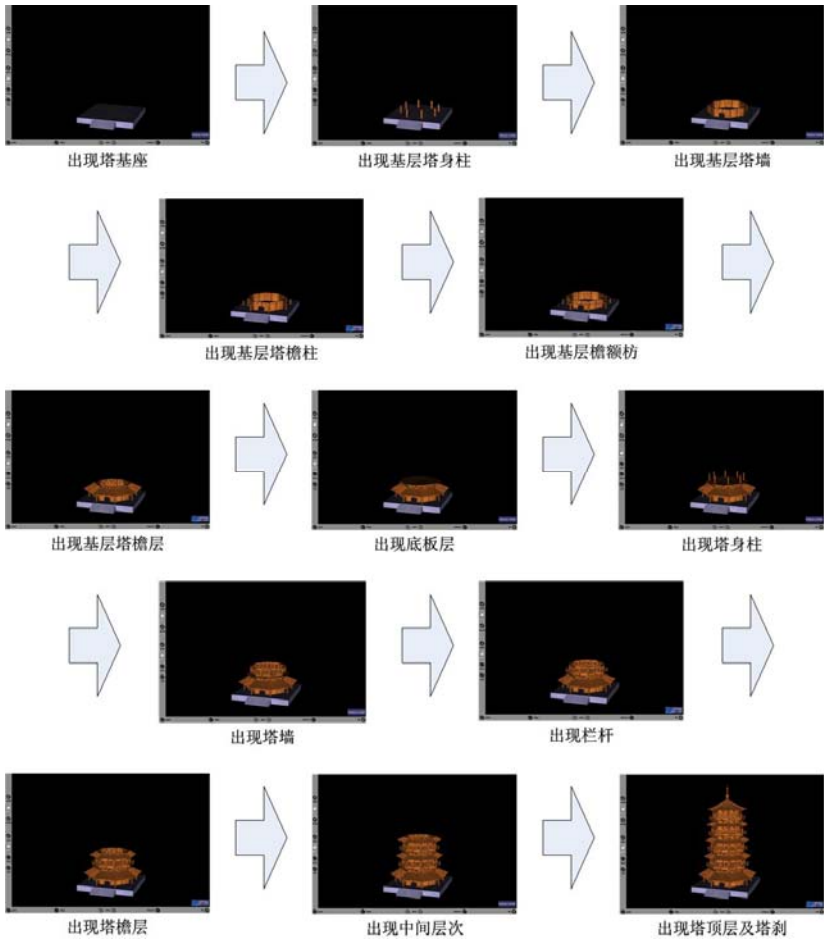


图 10 建筑生成动画展示 (VRML 版本)

4 总结

《古建筑动画系统》中的《塔类建筑平台》集成了如楼阁式塔、密檐式塔、覆钵式塔、金刚宝座塔、花塔和亭阁式塔等六种在中国国内较为常见的六种塔类建筑。系统具有良好的扩充性，系统使用者可以通过在部件库中添加新的部件类型，在规则库中添加新的依据营造法式总结得到的构建规则来向系统中添加新的建筑类型。因为现实中的塔类建筑缺乏固定的法式，造型千变万化，因此对现实中存在的具体的塔类建筑的表现可以通过修改中间文件实现。

但《塔类建筑平台》也存在一些不足，尽管提供了良好的可扩充性，但是因为塔类建筑部件库在设计时，对塔类建筑的部件划分的粒度较粗，即每一个部件包含的内容较多，因此如果在最小的粒度层次中的部件存在细节上的差异，那么除非在部件库中添加新的部件，否则难以通过已有的部件对这种细节上的差异进行体现。

5 致谢

感谢刘椿年老师和张松懋老师对于本项目及本篇论文的悉心指导；同时感谢孔亮、孙嘉、顾博、王妍、冯佳、张昊、宫丽环、谢明皓、孙凯，王巍峰，未宫瑾，尹梅芳、白卫静等同学为本系统的实习提供的无私帮助。

参考文献

- [1] Raibert M Hetal. Animation of Dynamic Legged Locomotion. Computer Graphics[J]. 1991, 25(4):349-358.
- [2] 夏利民，古士文，沈新权. 基于水平集的 3D 动画[J]. 计算机研究与发展. 2002, 39(2):236-241.
- [3] Lu Ruqian, Zhang Songmao. Automatic Generation of Computer Animation[J]. Springer-Verlag, 2002:33-35.
- [4] 陆汝钤，张松懋. 从故事到动画片-全过程计算机辅助动画片自动生成[J]. 自动化学报. 2002, 28(15):321-348.
- [5] 罗哲文. 中国名胜——寺塔桥亭[M]. 北京：机械工业出版社，2006.
- [6] 罗哲文，刘文渊，刘春英. 中国名塔[M]. 天津：百花文艺出版社，2006.
- [7] 孙建华. 漫步古塔名楼[M]. 北京：中国社会科学出版社，2005.

作者简介

刘射彪，男，1985 年，北京工业大学硕士研究生，河北省石家庄市，研究方向：人工智能；
梁天柱，男，1980 年，北京工业大学硕士研究生，吉林省延吉市，研究方向：人工智能。

高校办公室知识管理维度分析

陈学东 刘文娟

(北京交通大学经济管理学院, 北京 100044)

摘 要: 人和技术是知识管理的两个重要维度, 高校办公室知识管理策略就是要在这两个维度上做出选择。基于此, 本文提出了高校办公室实施知识管理的信息化、人性化和综合化三种策略。本文从高校办公室知识管理的内涵引入, 提出了高校办公室知识管理的两个重要维度, 并且提出了相应的策略, 在此基础上, 对高校办公室知识管理的激励机制进行了简要的概述。

关键词: 高校办公室; 知识管理; 维度

The Dimension Analysis of Knowledge Management in University Office

CHEN Xue-dong LIU Wen-juan

Abstract: Human and technology are two important dimensions of knowledge management, the strategy of knowledge management in University Office is to make a choice on these two dimensions. Based on this, this paper presents three kinds of strategies just as information, human, integration strategies in University Office when it implements the knowledge management. In this paper, first, introduce the connotation of knowledge management in Universities Office. Second, raise two important dimensions of knowledge management in university offices, and propose a corresponding strategy.

Keywords: university office; knowledge management; dimensions

1 高校办公室知识管理的内涵

20 世纪 90 年代以来, 知识管理的理论与实践逐渐丰富起来。美国的卡尔·弗拉保罗认为“知识管理就是运用集体的智慧提高应变和创新能力”, 是为实现显性知识和隐性知识共享提供的新途径; 斯维拜从认识论的角度对知识管理进行了定义, 认为知识管理是“利用组织的无形资产创造价值的艺术”; 阿比克将知识管理活动定义为对组织知识的识别、获取、开发、分解、使用和存储。高校办公室实施知识管理是高校在知识经济时代的一次管理革命, 是积极应对知识经济挑战的一种超前战略选择。根据人们对知识管理的不同理解, 从高校办公室的性质和管理的本质考虑, 本文认为, 高校办公室知识管理是在充分肯定知识对高校办公室价值的基础上, 通过创造一种环境让每位办公室人员能获取、共享、使用组织内部和外部的相关知识信息以形成个人知识, 并支持、鼓励个人将知识应用、整合到服务中去的一种全新

的管理^[1]。

2 高校办公室知识管理的维度

在知识创造与传播过程及知识管理活动中，有两个因素至关重要：一是人，二是技术。这两个因素同时构成高校办公室知识管理的两个维度。

人之所以是知识管理的关键因素之一，是因为人（的大脑）不仅是隐性知识的载体，而且是知识创造和传播的内生力量，其具备沟通、团队合作、学习、应变、自我发展、协调、创新、自我控制、自觉等多方面能力。在知识创造和传播的四个阶段，每一个阶段都离不开人的参与，几乎完全是人的因素在起作用。办公室人员通过师徒制、同事交流、岗位轮换等方式传播隐性知识，其中观察、模仿和亲身实践起决定性的作用。办公室人员通过团体工作、在职培训、学院教育消化和吸收新获取的显性知识，进而创造新的隐性知识。因此可以说，人是知识创造与传播的决定性因素，也是知识管理的重要维度之一。

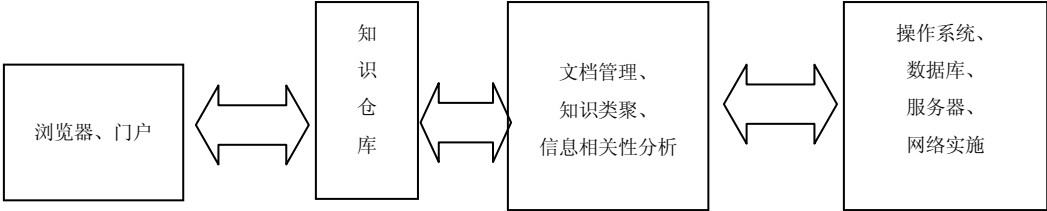


图 1 知识创造和传播的四个阶段

技术主要是在知识创造与传播过程中的组合阶段起作用，同时也支持外在化与内化过程。在外在化阶段，电视会议系统、电话、E-mail 等通讯和信息技术能够强化和方便人们的沟通和交流，因而也促进了隐性知识向显性知识的转化过程。在内化阶段，计算机仿真、虚拟现实等技术可以向人们提供实时的（just in time）培训，MIT 组织学习中心开发的微世界（microworld）^[2]就是这方面生动的例子。因此，技术在知识创造与传播过程中也起着关键作用，是知识管理的重要维度之一。当然，与人相比，技术只不过是一种使用工具，并不能成为知识管理的内生力量。

组织可以通过创建适宜的组织环境和加大在信息技术方面的投资力度来强化知识管理过程中两个维度的作用。项目团队、特别任务组（work forces）等正式团体由于其良好的沟通性能，被西方学者认为是组织中最佳的学习单元^[3]，因而也被西方企业广泛采用。这种正式的工作团体鼓励面对面的交流，促进知识创造与传播过程中社会化和内化两个阶段的知识转化与吸收，因而在知识的创造与传播过程中起着重要作用。近年来，西方企业在积极完善正式工作团体的同时，又开始着力培育象“实践社团”（communities-of-practices）^[4]这样的非正式团体，使正式团体和非正式团体成为组织中两个互为补充的知识创造与传播系统。非正式团体成员来自相同的专业领域，使用相同的专业术语，因而更容易交流，可以促进外在化过程，同时也有助于社会化和内化过程。有的学者调查后发现，员工在工作场所获取的知识中，有 70% 来自于非正式团体成员的交流和沟通^[5]。因此，创建正式的工作团体，培育非正式的学习团体，使二者互为补充，是知识管理过程中发挥“人”的因素的组织基础。信息技术不仅支持显性

知识的快速存取，而且支持人与人之间的快速沟通，因而也支持知识管理过程中“人”的因素的发挥。不少公司投入巨资建设基于知识的系统。在这样的系统中，知识库可供人们存取编码化的显性知识，知识地图可供人们寻找尚未编码、仍储存于人们头脑中的隐性知识。此外，还有 E-mail 系统、电子图书馆、网上论坛和虚拟会议室等。

激烈的社会竞争从深度和广度上推动了信息技术的应用，它以网络技术、计算机技术和信息技术为基础，帮助高校办公室对相关的知识资源进行明晰化、系统化的管理。它包括建立团队协作的专家网络，让所有人都能快速而方便地访问到或学习到所需要的信息和知识，无论数据库、文档、政策、业务流程还是办公室人员头脑中的知识和经验，都能够得到高效的共享、利用，使恰当的知识在恰当的时间通过恰当的场合和载体传递给合适的人，从而使高校办公室向数字化、网络化、知识化、虚拟化和全球化方向发展。因此，技术在知识创造与传播过程中也起着关键作用，是知识管理的重要维度之一。当然，与人相比，技术只不过是一种使用工具，并不能成为知识管理的内生力量。

组织可以通过创建适宜的组织环境和加大在信息技术方面的投资力度来强化知识管理过程中两个维度的作用。

3 高校办公室知识管理的策略

本文以知识管理的两个维度为出发点，将高校办公室的知识管理策略分为三种：信息化策略、人性化策略和综合化策略（见图 2）。

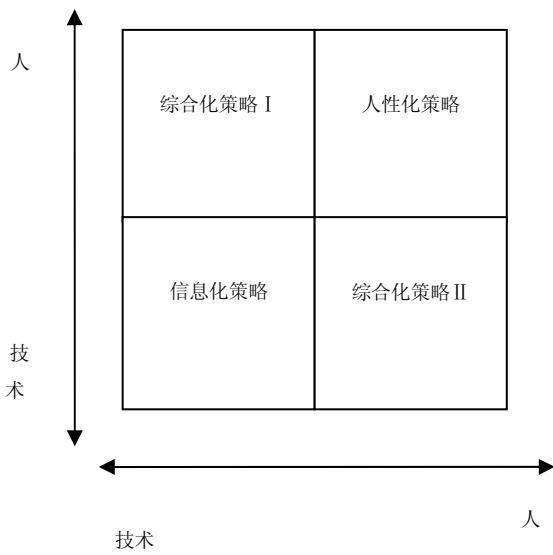


图 2 高校办公室知识管理策略

信息化策略是指单从技术一个维度管理知识的策略。此类高校办公室的日常活动主要依靠显性知识，很少依靠员工头脑中的隐性知识。那些一成不变地处理事务，千篇一律地对待问题的高校办公室，大多采用信息化策略。由于这类组织主要利用原有的知识进行重复性的工作活动，因此快速获取知识是高校办公室制胜的关键，而信息化策略正好可以满足这一要

求。该类办公室主要需求成熟的、标准化的计算机网络系统或知识管理系统方案,使用的主要是显性知识,因此高校办公室内部的知识管理系统对该类办公室的实施非常重要。可采用“从人到文档”(people-to-documents)的方法将员工所拥有的隐性知识显性化,从而使全体员工能够最大限度地利用办公室的知识管理系统创造价值,因此也可获得巨大的成功。或者其基本战略可确定为“利用原有的知识而不是创造新知识来工作”,可以开发一个专家系统,将日常工作种可能出现的问题和专家建议的解决方案输入系统,办公室只雇佣那些资历较浅的工作人员就可以顺利完成工作,员工甚至通过电话就能得到专家系统的指导,可大大降低办公室的成本。例如高校办公室中的财务部门就可以运用相对标准的系统实现繁琐的工作,简化工作量,节约时间与成本^[6]。

人性化策略是指单从“人”一个维度管理知识的策略。这类高校办公室的日常活动主要依靠员工头脑中的隐性知识,而不是办公室现存的显性知识,办公室的基本战略也是以创新多变的技能而不是一成不变的技能维持日常工作,高校办公室更多地需要多变的而不是标准化的服务,如高校办公室中处理学生助学贷款等部门,它的基本职能的实现主要依靠办公室员工的隐性知识而不是办公室现存的显性知识,这个办公室的知识管理策略就是一种“从人到人”(people-to-people)的人性化策略。工作人员与学生之间的知识传递主要依靠面对面的交流,通过“讲故事”(storytelling),而不是计算机网络。这个办公室进而致力于在工作人员之间培育非正式学习团体,使办公室的工作开展获得了成功,避免了因信息不对称而存在的执行过程中的遗漏与错误。

信息化策略与人性化策略都是在一个维度上管理高校办公室的知识,实际上对于同一所高校的各个办公室之间是在两个维度上管理自己的知识,我们把这类高校办公室的知识管理策略叫做综合化策略。根据在两个维度上的侧重点不同,综合化策略又可以分为以人为主的综合化策略和以技术为主的综合化策略。以人为主的综合化策略强调人在知识管理过程中的作用,知识管理的手段以“从人到人”的共享隐性知识模式为主,但同时又开发虚拟会议系统、网上论坛以及知识地图以方便和促进人与人之间的交流。在这种形式中,技术是作为一个辅助工具而不是一个主要因素。而以技术为主综合化策略则强调技术在知识管理过程中的作用,以“从人到文档”的隐性知识显性化模式为主,技术是该策略的主要因素,而人只不过是辅助因素而已。人的作用就是将自己头脑中的隐性知识用专业语言表达出来,从而能够输入计算机系统供其他人分享。实施综合化策略比较成功的高校办公室都是那些在两个维度上有侧重点的办公室,而不是两者同时并进。

4 高校办公室知识管理的激励机制

通过各种有效的激励手段,激发人的需要、动机、欲望,形成某一特定目标并在追求这一目标的过程中保持高昂的情绪和持续的积极状态,是人性化管理的重要机制。

首先,激励的基本原则是尊重。其中包括尊重员工的生命价值,尊重员工多样的兴趣爱好和生活方式,激励员工的思想自由,营造宽松环境,发挥员工自身的能动性和创造性。要尊重员工的劳动成果,鼓励他们努力工作,积极向上,奋斗不息,使其个人目标的实现与组织目标落实紧密地结合起来。其次,奖惩也是人性化管理的基本手段。在知识管理中,坚持奖惩结合、以奖为主的原则,公平合理,根据不同的人和人的优势进行奖励,以精神奖励为

主。再次,竞争是提高激励效应的推动力。如有些高校办公室已实行的竞争上岗,由于工作任务的挑战性,促使了员工由被迫学习转变为要求学习、主动学习、自觉学习。这说明只有竞争才能激发人的进取心、主动性冒险精神和创造性思维,而这些正是知识创新和技术创新所必需的心理特质。参与也是激励的一种重要方式,参与是管理者通过一定的制度和形式,让员工参与组织的决策、计划的制定、对某些事情的处理和对某些问题的讨论和管理。作为知识型高校办公室的员工,多数都具有较高的科学文化水平和管理能力,民主意识和参与欲望都很强。通过参与能使员工与办公室相互依存、和谐发展,促进个人知识与组织知识的有效转化。

5 结语

人类正在跨入知识经济时代,管理知识成为高校办公室管理的主要内容,实施什么样的知识管理策略对于高校办公室日常工作的顺利展开至关重要。本文提出人和技术是高校办公室知识管理的两个主要维度,并由此将高校办公室的知识管理策略划分信息化、人性化和综合化三种。高效办公的知识管理策略由日常工作的简易程度,创新性程度决定。那些使用已有的显性知识为主要战略,日常工作标准化程度较高而创新性程度较低的办公室,应当采取信息化策略,或以信息化为主的综合化策略;而那些以使用隐性知识为主要战略,日常工作以定制化为主或者创新性程度较高的办公室,则应当采用人性化策略,或以人性化为主的综合化策略。正确选择知识管理策略往往是高校办公室顺利开展工作的有效保障。

参考文献

- [1] 彼得·圣吉.第五项修炼——学习型组织的艺术与实务[M].上海:上海三联书店,1998.
- [2] 宋庆红.加强高校知识管理提升其竞争力[J].理工高教研.2004年第6期.
- [3] 彼德·德鲁克.知识管理[M].北京:中国人民大学出版社.1999.
- [4] 陈玉强.学校办公室知识管理模式的构建研究[J].科学教育家.2008年6月第6期
- [5] 孙雪松.试谈高校办公室知识管理模式的构建[J].北京联合大学学报(自然科学版).2004年12月第18卷第4期总58期.
- [6] James M. Bloodgood. Understanding the influence of organizational change strategies on information technology and knowledge management strategies. Decision Support Systems 31 (2001) 55-69

作者简介

陈学东,男,1972年10月生,内蒙古人,北京交通大学经济管理学院副教授,研究领域:企业信息化;知识管理。

刘文娟,女,1987年2月生,研究生,研究领域:知识管理;智能决策支持系统。

Return to the Origin of Architectural Design: Based on the Research and Application of Sketchup in the Teaching of Architectural Design

TANG Hong WANG Jin-yu

(Henan University of Urban construction, Henan Pingdingshan 467000)

Abstract: This passage briefly introduces the application features of the software-Sketchup which can be used in the practice of architectural teaching. And the students are guided to pay much attention to the architectural form and the space in design so as to thoroughly understand the origin of architectural design, in this way the students' professionalism will be enhanced.

Keywords: Sketchup; architectural design; architectural form; architectural space

1 Instruction

Mix clay to form the implement, just because it is hollow, it is useful. Chisel the door and the windows to build a house, just because it is hollow, it is useful. So the "exist" could boring us facility; the "empty" play an important role^[1]. More than 2500 years ago, Laozi propose that the value of the architecture is the space which it formed. the modern architectural concept emphasize that the space is the essence of architecture, it require us to study and comprehend the architecture in it's space and the form so that we can carry on the architectural creation to form a architecture with graceful feather and suitable space. At the same time, architecture education emphasize to culture the students the ability of space shaping. However, due to various reasons, architectural education now has deviated from this subject.

2 lack of Architectural Education at present

1) The plight of Architectural Education during the Information Age

During the 60-70 years of the 20th century, we has been step from industrial society into information society, this means the communication has been fundamentally changed. Digital and information-based means of communication has been infiltrated into various fields. Architectural design industry is of one of the industries who early realization of computer-aided design. As technology advances, there is essentially difference between the modern architectural design patterns and the traditional design patterns. Architects can use compucture to build architecture digital model, directly carry on the architecture design through three-dimensional views without spending to much

energy for space conversion^[2]. With the information technology penetrating deeply, architectural design is moving towards a paperless era of digital design. But the current architectural education still cling to the traditional teaching methods, the talent is unable to adapt to today's design environment.

2) Ignore the architectural design process

Regarding a specific program, it is better for the teacher not to comment the program too detailed, otherwise it easily to restrict student's cogitation, and the students will make a lot of identical design. It is hard for the students to design if they didn't get specifically guide from the teachers: At first they often do the plan functional division according to the architecture features and the task book, and then they consider the introduction of elevation and section when the plan have been carried to a certain extent. It is hard for the students verbally to express their design ideas clearly just by their several sketch, and the teacher can hardly to understand the students' creative intents^[3]. Though the teacher always communicate with the students, they often neglect the architect design progress, it is hard to enhance the students' design ability.

3) Ignore the architectural environment

In reality, each architects have their specific existents, and there is some relationship between the surrounding environment and the architecture. Most students never consider the surrounding environment carefully to make the architecture coordinate with the environment, and never consider the how the surroundings restrict the architecture programs. Many students never do a wide range of research about the surroundings environment when they get the task book, their knowledge about the site is just limited to the task book. Because the information the students collected are not enough, they hardly thinking comprehensively when they design. The design which the students made out often full of loopholes or became the castles in the air and hardly to built.

4) The students lack the knowledge about the construction material, and didn't understand the architectural detail enough

Architecture are composed of different materials, and materials are the emotion symbols of the architecture. There will have different effects if we endow the same architecture with different materials, even the effects are extremely opposite. A qualified architect can skilled use of various materials. Unfortunately, now the students didn't aware of the importance of architectural materials yet. The architectural details are the important aspect to reflect the professionalism of the architects. Skilled architects are better in the construction details design. Now the students' programs hardly involve the two aspects.

3 Sketchup

1) R & D background

At present, though the computer technology are widely used in the design of construction industry, but this application is only made the software as a drawing tool. Architectural design software and not play right role in design, it just demonstrates the superiority at the end of the

program. There are two reasons to cause this effect: one reason is the software itself. Before the Sketchup, a large number of architectural design software can not to adapt to the architect's work habits. The line in the two-dimensional design software (such as AutoCAD) always restraint the architect's thinking, 3D Design Software (such as 3dsmax) lack the ability to control two-dimensional graphics. Another reason is that the architect is not really aware of the meaning and advantages of computer-aided design. During the design they adopt the traditional design process and add a computer plot course to instead the original hand-drawing process.

2) The features of Sketchup

At beginning, Sketchup are made to “A specifically medium in order to explore motif and synthetic informations”^[4]. This software have concise user interface, operate easily, powerful function. (figure 1). The most important is that it conforms to the architect's work habits, completely catered to the architect's work plans. There is little operation system in Sketchup, the novice could just familiar with a few commonly used commands and then designing the program in the Sketchup as the same as they Sketching with a pencil. Architects can quickly use Sketchup to manifested their own design inspiration in the form of three-dimensional ways. It is easily to revise the design express them in many ways. Sketchup communicate perfectly with the revelatory of the architects' sketch design and the certainty of the computer software design, it is not a conditional design software but a real architecture design software.

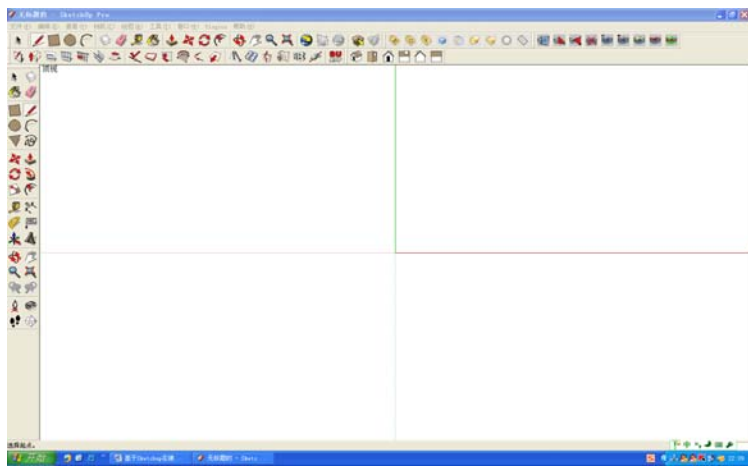


Figure 1 concise user interface of Sketchup

4 The use of Sketchup in architecture education

Aiming at the shortcomings of modern architecture education, It have enormous practical significance to bring Sketchup into architecture education according the software's characteristics, they can enhance the effectiveness of architecture education to a large extent and help the students step out of the architectural design errors to the origin of architectural design.

1) The meaning of architectural lies in its three-dimensional space, The main content of

architectural design is the architectural space and the physical of the construction

Sketchup itself is a three-dimensional design software having a very strong three-dimensional design functions.(figure 2).In Sketchup, the architecture design is carried on the three-dimensional space from the initial concept module body research to the finalization of the construction program. Architects can observe each step of modification from the three-dimensional through multi-angle and then to determine whether the modification is reasonable. But the conditional design process are carried on the two-dimensional environment. The architect have always to consider what the three-dimensional space looks like, when they do the three-dimensional model, they have to consider whether it consistent to the two-dimensional design. Thus architects design process appears tortuous. The aim is to design three-dimensional space and the architecture form, but had to be achieved by means of two-dimensional means, the architect waste a lot of energy. If we use

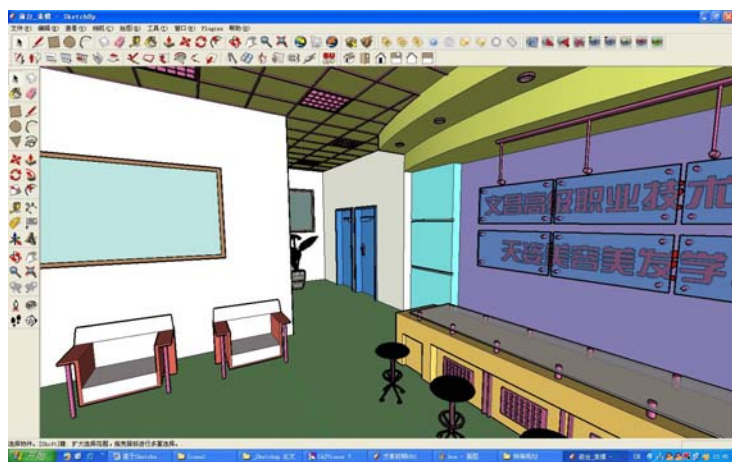


Figure 2 three-dimensional space

Sketchup, we can omitted the conversion between the three-dimensional design and the two-dimensional design process, it is true three-dimensional design and the architects will have more energy to pay attention to architectural design itself.

During the process of architecture education, Sketchup lifted restrictions of two-dimensional design, let the students never restrict by the two-dimensional design. they can do the architectural design at higher and more direct point of view. Sketchup can also help the students enhance the ability to control architecture space and architecture form. After we use convenient and efficient means built model by Sketchup, we can observe, amend, or compare the architecture form, interior space, architectural detail in various angles, thereby we can constantly improve the design and to enable students to get to know the origin of architectural design, we can enable students to understand the real architecture design is not just the plans, elevations, sections and the unrealistic effect pictures getting together, but a architecture three-dimensional shape design. In addition, through a round of changes can let the students understand the construction program design is a gradually process, To establish a program can not be achieved by just one design, during this period, the design must be constant change and improved.

After completed the model in Sketchup, we can rapid amend the model using groups and components of its functions, this can help the students to capture architectural design inspiration. And contrast through a multi-group model is conducive to enhancing the students building physical analysis, Sketchup provides slice surface features which provides a convenient way to analysis architectural interior space. By this function can clearly observe each direction of the indoor space of the construction.

2) Expression of the construction program

As a specifically construction software, Sketchup emphasize on the building program expression: (1) Seeing shall be obtained. without a long wait for rendering, architectural models can be observed directly. (2) Multi-angle view and multi-mode instrument. Sketchup offer us the perspective, top view, back view, left view, right view five kinds of observation angle, it is easy to observe various elevations and perspective of the architectural model. And can use mouse drag, rotate, zoom in, and then we can observe the architectural model from any point of view. Additionally, Sketchup offer us X-ray model, wireframe mode, blanking mode, color mode, map mode, monochrome mode six kinds of display mode for different architecture design performance. (3) The use of the page, regarding some important elevation or the perspective, Sketchup can set free page. Through these different point of view and the different pages of display modes we can analysis the focus point of view of the architecture. (figure 3,4). (4) The phased and diversification technique of expression can meet the needs of design expression. Sketchup can construct kinds of Three-dimensional model according to the performance characteristics of different stages of architectural design, some of them are concept sketches, some of them are delightful and full of artistic feeling, others are plain, and actually obey the virtual design effect. Therefore, the architect can express the design objects in accordance with the design phases, and then to provide the owners with the performance results relevantly^[5].

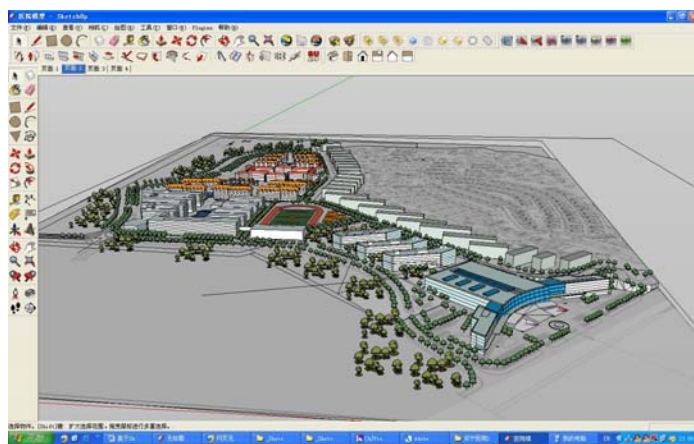


Figure 3 The use of the page

During the communication of the architectural design, architects can use Sketchup or other architectural design software to create a virtual scene to express design intents. This virtual building can deepen the synergy between the design of relevant person^[6]. Teachers and the students can

communicate directly during the architectural education by using a variety forms of Sketchup copy expression. Students can express the teachers their own design ideas and the development of the middle design process or the final establishment program.The teachers can observe the students' design from Multi-angle in Sketchup and then to understand the students design intent, clearly the process of program in-depth, analysis and establish program's strengths and weaknesses,and then to guide and help the students during the whole design process.

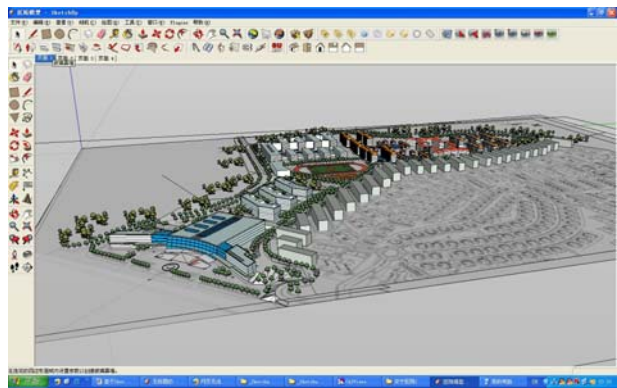


Figure 4 The use of the page

3) To cultivate the students' environmental awareness during the architectural education

During the Sketchup application,we can construct the three-dimensional model space environment of sites by simple and convenient modeling tool. During the three-dimensional environments,teachers can observe the architecutral program at any angle,and then help the students develop the design.(figure 5).At the same time,we can use the modue store of Sketchup to insert the components of the built environment according to the actual size of the object model,such as buildings, vehicles, characters, landscape,from this,we can learn more real sense of scale and the coordination with the surrounding environment.

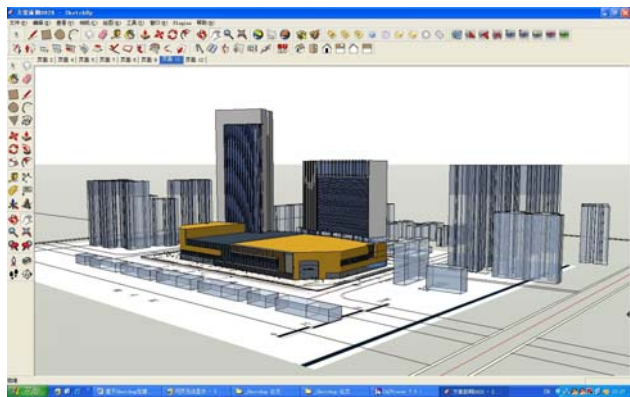


Figure 5 Environment Simulation

Additionally, the Sunshine Analysis System of Sketchup can analog sunset, sunrise, sunshine conditions, as solar altitude angle and so on. Architects can set a specific date of the analysis of sunshine and then drag the time slider to carry out the construction of the dynamic analysis of

sunshine, through this we can see intuitively the relationship between the building block and every building of the sunshine state (figure 6) .

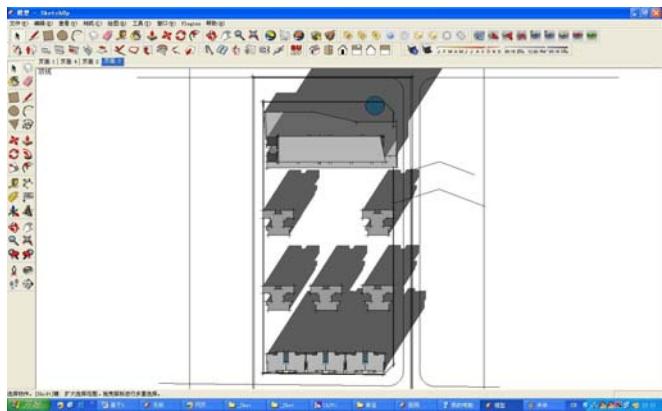


Figure 6 Sunlight Analysis

4) The building colors and materials applications

Building colors and materials design is a important stage in architectural design, but also is a relatively weak link in architecture education. The Sketchup material texture, color editing capabilities and the convenient means of material endue can help the students to to complete the deployment of architectural colors and materials easily(figure 7). Different from the 3D Model (Which must be rendered after the rendering device in order to observe the effects of different materials), in Sketchup,we can observe the modle immediately after given material, it is easy to observe the effect and adjust the material,and ultimately achieving the best visual effect.We can comparative analysis the different colors and materials of the architecture,or simulate how the environment effect architecture. This man-machine interaction is a new way which the computer software technology create for architects to analysis new ideas [7]. Through this approach, we can gradually culture students' ability to use colors and materials, and then culture students the architect's basic literacy.

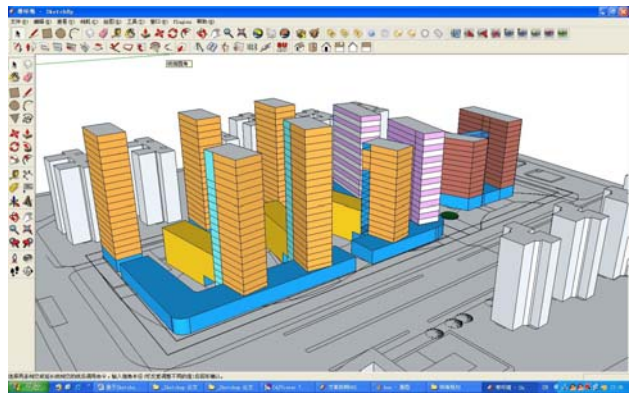


Figure 7 color Simulation

5 Epilogue

Sketchup can generate a program quickly, its flexible to amend a program and have convenience multi-angle expression, these demonstrate that it is designed to meet the design process and developed ,it can conform to the architect's work habits .these two advantage made it to be a true sense of the architectural design software. If we introduce Sketchup into the teaching practice of architecture, coupled with the correct guidance of the teachers can help students out of the current study and design errors, and truly understand the nature of architectural design.But Sketchup only is a tool-aided design, and can not completely replace the traditional teaching methods to complete the architectural design of the teaching tasks independently. Computer does can be seen as an extension of our hands, but dofferent from pencil, pen or typewriter ,it is a completely different medium. For this medium,we have to use new way to work^[8]. This requires us to combinat the the traditional teaching methods and the Sketchup-aided design methods perfectly in architecture education in the future, and will greatly improve the efficiency of the architecture education.

References

- [1] Li Yiran.Tao Te Ching [M].Beijing: China Radio and Television Press, 2007
- [2] Fu You. Transformation of Design Thinking and Method Caused by Building Information Modeling[J]. ARCHITECTURAL JOURNAL2009 (1): 77-80
- [3] Chen Qiuguang.The whole fragment-- -the research and practice of setting of the Introduction to architectural design curriculum [J].new architecture2009(5): 58-60
- [4] Li Jiancheng.Digital Architectural Design[M] . Beijing: China Building Industry Press.2007.62
- [5] Xiu Weiguo.Digital Construction[J].ARCHITECTURAL JOURNAL2009 (1): 61-63
- [6] BARBOSACAM,DREUXM,BENTOJ,etal.An object model for collaborative CAD environments[A].The 7th International conference on Computer Supported Cooperative Work in Design.NRC Research Press,2002:179-184.
- [7] Qu Jianmin. The Tutorial of Multimedia Multimedia Technology and Application[M]. Beijing: Tsinghua University Press.2005,4:24-25
- [8] Mark Lauden.The Architects Guide to Computer Aided Design[M].ACM Press.2004:24

作者简介

唐红,男,1970年9月生,硕士,河南潢川人,河南城建学院城市规划与建筑系,副教授、工程师、国家注册建筑师,建筑设计教研室主任。研究领域:建筑设计及其理论。

自主水下机器人能源系统设计

滕学志 魏志强 殷波 董艳

(中国海洋大学 信息科学与工程学院, 山东 青岛, 266100)

摘要: 能源系统是自主水下机器人重要组成部分, 电源管理的质量将会直接影响到系统的性能和硬件的使用寿命。文章提出了一种适用于自主水下机器人的电源管理设计, 采用单片机及外围电路实时的对机器人中设备的电压、电流和温度等信息进行全程监控, 出现故障时可以迅速做出相应处理。

关键词: 自主水下机器人; 电源管理; 单片机

The Design of Energy System for AUV

TENG Xue-zhi WEI Zhi-qiang YIN Bo DONG Yan

(College of Information Science and Engineering, Ocean University of China,
Qingdao Shandong 266100, China)

Abstract: Energy system is one important part of AUV. The quality power management system directly influences system performance and hardware's service life. This paper presents an power management design for AUV, monitoring the robot device voltage, current and temperature information by real-time MCU and peripheral circuits. When the system has failure, it can process quickly.

Keyword: AUV; Power management; MCU

1 引言

进入 21 世纪以来, 随着世界经济和军事发展的需求, 海洋资源开发、海洋能源利用等现代海洋高新技术的研究已成为世界新科技革命的主要领域之一, 其中深海资源探测与水下作业关键技术与装备已成为各海洋大国不遗余力进行研究的主要对象。自主水下机器人 (Autonomous Underwater Vehicle: AUV) 作为水下探测的主要工具, 在海底地形地貌探测、海洋工程建设、海洋资源开发、海洋科学探索以及维护国家海洋权益等诸多方面发挥极其重要的

基金项目:

(1) 国家 863 目标导向课题: 基于声纳和水下视觉的深海复杂环境下 AUV 组合导航系统关键技术 (2009AA12Z330)。

(2) 教育部博士点新教师基金: 基于多传感器数据融合的深海机器人 SLAM 自主导航方法研究 (20090132120013)。

作用。对于 AUV 而言能源系统作为保障其满足长程、深海、复杂环境下探测的重要组成部分，值得深入研究。

2 AUV能源系统分析

2.1 电池的选择

电源是水下机器人的重要组成部分。大部分的 AUV 都是以电力作为主要的驱动方式。以电力驱动为主要形式的 AUV，电池作为 AUV 的能量供应中的重要组成部分，它的选择需要多方面的考虑。除了考虑能量密度和能量/体积比，还应充分考虑使用效率、环境温度、噪声、充放电有无气体泄漏、水下密闭环境安全性能、对环境有无污染等因素。远程 AUV 可能采用的几种电池性能指标如表 1 所示^[1,2]。

表 1 化学电池及其性能指标

主要参数	能量密度（Wh/Kg）	体积比能（Wh/L）	充放电寿命（T）	自放电率（%/M）	污染
镍镉电池	50	150	500	25-30	有
镍氢电池	65	200	500	30-35	无
锂离子电池	105-140	300	1000	6-9	无
银锌电池	75-80	150-170	>100 次	10-11	无

综合各方面的因素这里选用锂离子电池装备成的电池组。对于水下机器人项目来说其他可能的选择或者太昂贵，或者本身就是重要的研究项目。此外，核能源涉及环境和法律问题，使它们无法得到大多数群体的使用。

本系统采用额定电压 48V 的 2 组锂离子电池组供电，一组作为正常工作时使用，另一组作为主电池出现异常、电量耗尽等紧急情况的备用电池使用。

2.2 AUV组成部件电压分配

为构建深海复杂环境下的 AUV 精确组合导航定位系统，这里采用多传感器装置，系统的主要能源分配如表 2 所示。通过 DC/DC 模块实现 48V 到 24V，与 48V 到 12V 的转换，12V 以下的低功耗装置采取由 12V 接三端稳压器件直接转换得到。

表 2 主要能源分配情况

设备名称	工作电压（V）	数量（个）	实现功能
直流无刷电机	48	5	实现 AUV 六自由度运动
结构光发生装置	24	1	实现基于结构光的 AUV 中尺度自主导航
工控机	24	1	中央处理机，综合各传感器信息利用 SLAM 算法实现复杂海洋环境下 AUV 的组合精确自主导航
扫描成像声纳	24	2	用于 AUV 的障碍物和环境探测，实现大范围复杂环境下的 AUV 大尺度全局自主导航
深度计	24	1	基础导航组件提供 AUV 深度信息
底层控制系统	12	1	底层运动驱动、传感器数据采集、及转换
GPS	12	1	基础导航组件提供 AUV 下水前位置
航姿参考系统	6	1	基础导航组件提供 AUV 航姿信息
数字罗盘	6	1	基础导航组件提供 AUV 航姿信息
双目视觉平台	5	1	用于近距离的环境和障碍物探测，实现小尺度的 AUV 精确自主导航

3 电源管理系统主要功能

电源管理系统通过采集各个耗电模块的电压、电流、和电池及测试点温度等信息通过计算得出系统剩余工作时间的估计值和系统的工作状态，通过 RS232 接口与负责 AUV 决策处理的中央工控机进行通信，通过对每一路情况的分析，为中央控制系统提供决策，进而对每一路分别进行相应的供电控制。在出现问题时应急处理模块启动，对故障进行处理保障 AUV 顺利返航。电源管理系统结构如图 1 所示。

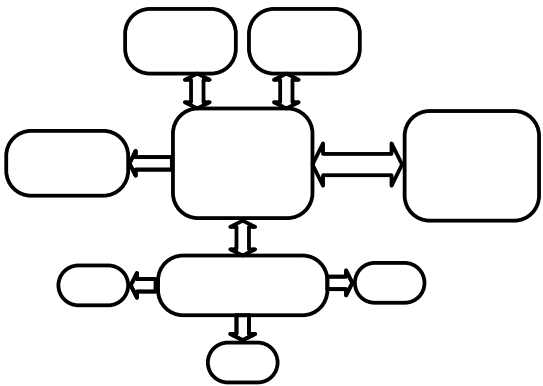


图 1 电源管理系统结构图

系统分为四大功能模块：数据采集模块、开关控制模块、电量计算模块、应急处理模块。数据采集模块负责对电源输出的电压、电流、电池组与 AUV 工作仓内电压等周围的温度等模拟量进行采集和预处理，为系统状态分析提供依据。

开关控制模块通过驱动电路控制继电器的开关，从而对每一个耗电设备的电力供应进行管理，同时便于在某一设备出现短路等故障时，不影响其他设备的供电。

电量计算模块采用每秒平均电流值作为当前单位时间的电量进行累加。存储下来后，根据电池总电量、已消耗电量和当前工作电流，可以计算出系统所能工作的剩余时间，可为整个系统的运行提供参考^[3]。

应急处理模块在电池电量不足时，结合电池组电压情况，通知工控机并及时切换备用电源，根据工控机指令关闭相应设备减少电源消耗，同时做好上浮准备。在总电流过流时，系统将关闭所有通路，并进入异常处理模式。在异常处理模式中，电源模块结合各测试点电流、温度值，快速查找出相应的故障通道，检测出并关闭故障通道，并向上级工控机报告状态。

电源管理系统从监控节点采集模拟量输入(如电压、电流、温度等)，进行电量计算同时通过 RS232 接口接收工控机的指令，对数据预处理，对开关量(如继电器通断状态)进行相应操作，在出现故障时调用应急处理模块。

4 电源管理系统硬件设计

本文控制器选择 ATMEL 公司生产 8 位单片机 AT89C52。P0 口负责各通路开关量控制，

P1 口负责电压、电流信号的采集，P2.0 口负责温度信号的采集。同时扩展 RS232 接口进行与工控机的通信。电源管理系统连接图如图 2 所示。

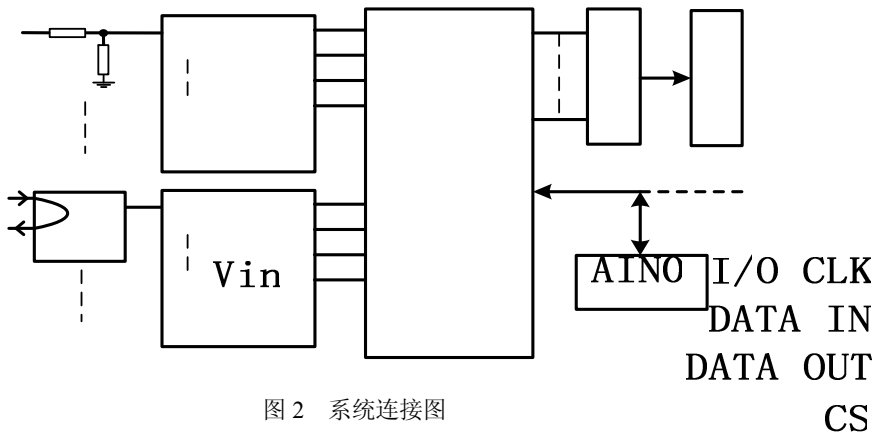


图 2 系统连接图

4.1 电压监测

电压信号的采集采用 TI 公司生产的 12 位 ADC 芯片 TLC2543，它具有 11 路模拟输入通道、10 μ s 的转换速度、片内系统时钟，采样精度达到 12 位，外部时钟最高可达 4.1MHz，能提供较高精度且多通道的数据采集功能^[4]。信号以串行方式输出只需要单片机 4 个引脚就可以对 11 路通道进行采集。各部分电压信号经过分压然后输入到 TLC2543 的模拟输入端，由单片机 P1.0-P1.3 输入。

4.2 电流监测

电流信号的采集采用 LEM 电流传感器 LTSR25-NP 测量实时电流。该元件是基于霍尔效应的闭环（带补偿的）多量程电流传感器，采用单极性电压供电，具有出色的精度、良好的线性度、无插入损耗、最佳的反应时间和电流过载能力^[5]。额定电流为 25A，最高可测 80A 的电流，满足系统设计的要求。该电流传感器可把电流转为 0-5V 的电压信号输出，然后通过 P1.4-P1.7 所连接的 TLC2543 转换成数字量，送至单片机。通过单位时间对电流量的累加，同时可得到电池所消耗的电量。

4.3 温度监测

温度检测采用 DALLAS 公司生产一线式数字温度传感器 DS18B20，测温范围-55 $^{\circ}$ C~+125 $^{\circ}$ C，固有测温分辨率 0.5 $^{\circ}$ C，测量结果以 9~12 位数字量方式串行传送。在使用中不需要任何外围元件，与微处理器连接时仅需要一条口线即可实现微处理器与 DS18B20 的双向通讯^[6]。非常适合于远距离多点温度检测系统中。这里将其用一根系统总线连接于单片机 P2.0 口。将其分布于电池组和其他大功率器件附近，实时的对整个 AUV 系统的温度进行监控，为电源管理系统分析系统故障提供依据。

5 电源管理系统软件设计

根据设计要求，电源管理系统的软件流程图如图 3 所示。

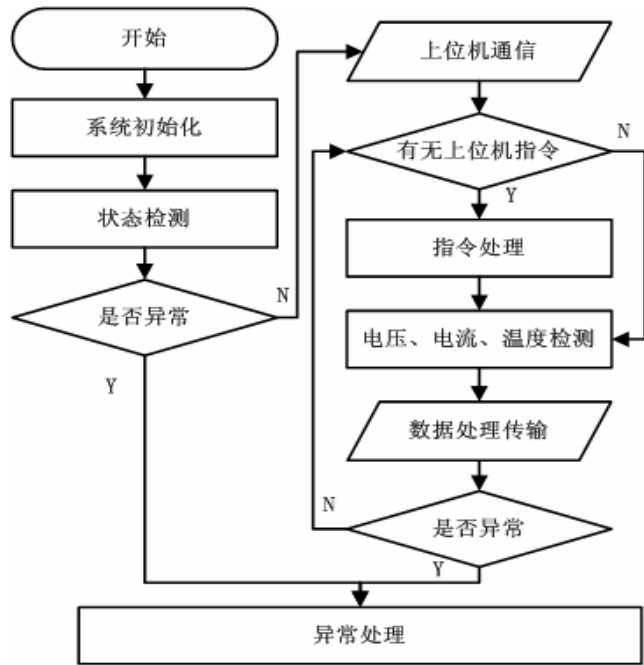


图 3 电源管理系统软件流程图

AUV 下水前，系统上电后首先进行系统的初始化，这时各个设备处于关闭的状态下，数据采集模块依次对各个节点的电压情况进行检查，然后对漏电流情况进行测量，如果出现异常超过设定的阈值，则调用异常处理子程序，找出故障点，启动蜂鸣器报警同时等待工控机决策。便于在 AUV 下水前及时发现潜在问题。

如果整个 AUV 电源系统正常，则下水后通过中断接收工控机的指令，根据工控机要求开启相应的设备，关闭长时间不用的设备，满足 AUV 整体低功耗的要求。并对各个测试点的电压、电流、温度情况进行实时监测，出现异常调用异常处理子程序。

6 结束语

本文探讨了一种 AUV 的电源管理系统的设计与实现。系统以单片机为核心，充分利用单片机的 I/O 资源，结合外围的传感器对 AUV 中设备的电压、电流和温度信息进行全程监控，并且在系统出现各种故障时可以及时查找并处理故障。本系统同时亦可用于其他类型和用途的需要电池供电的控制系统中，有着广泛的应用前景。

参考文献

- [1] Albert M. Bradley, Michael D. Feezor, Hanumant Singh, and F. Yates Sorrell. Power Systems for Autonomous Underwater Vehicles. IEEE Journal Of Oceanic Engineering, VOL. 26, NO. 4, October 2001.
- [2] 燕奎臣, 李一平, 袁学庆. 远程自治水下机器人研究[J]. 机器人. 2002, 24 (4):299-303
- [3] 刘丞, 赵建. 用于智能移动机器人的电源模块设计与实现[J]. 仪表技术. 2009, (2):67-70
- [4] Texas Instruments. TLC2543 - 12-Bit Analog-to-Digital Converters with Serial Control and 11 Analog Inputs Datasheet[Z]
- [5] LEM. LTSR25-NP - Current Transducer Datasheet[Z]
- [6] Dallas Semiconductor. DS18B20 - Programmable Resolution 1-Wire Digital Thermometer Datasheet[Z]

作者简介

滕学志 (1983—), 男, 山东青岛人, 中国海洋大学信息科学与工程学院电子系硕士研究生。主要研究方向为智能测控技术。

魏志强 (1969—), 男, 中国海洋大学计算机系教授, 博士生导师。主要研究方向为人工智能, 图形图像处理。

殷波 (1976—), 男, 中国海洋大学计算机系讲师, 博士。主要研究方向为机器人智能控制。

董艳 (1983—), 女, 中国海洋大学计算机系工程硕士。主要研究方向为软件工程与智能信息系统。

电子设计综合实验教学改革研究

王建新 李 莉 高献伟 路而红

(北京电子科技学院, 北京, 100070)

摘 要: 本文针对电子设计综合实验课程实践教学模式, 结合电子类课程教学的自身特点, 以激发学生学习兴趣, 培养学生创新能力为主旨, 从实验内容的设计、实验功能的实现、实验成绩的考核以及实验室的建设四个方面探讨了进行电子类综合实验教学改革的方法, 为构建合理的电子设计综合实验课程教学体系提供借鉴和帮助。

关键词: 电子设计; 综合实验; 教学改革

Electronic Design Comprehensive Experiment Teaching Research

WANG Jian-xin LI Li GAO Xian-wei LU Er-hong

(Beijing Electronic Science and Technology Institute, Beijing 100070, China)

Abstract: This paper studies the teaching mode of electronic design comprehensive experiment combined with the electronic teaching characteristics, in order to stimulate student learning interest, train the innovation ability. This paper discusses the teaching reform of electronic experiment from the content of experimental design, realization, examination and laboratory construction to construct a reasonable comprehensive electronic experimental teaching system.

Keywords: electronic design, comprehensive experiment, teaching reform

对于高校电子类课程教学而言, 实验是其中重要一环。当前, 大学生创新能力和综合素质的不足主要体现在知识的理解运用方面缺乏综合性实验环节, 课程设计、专业实习缺乏设计性实验内容, 致使实验教学水平低下, 教学模式僵化死板。电子电路仿真设计与实际电路制作是有一定差距的, 因此, 有必要将大学生的动手能力与知识综合运用能力结合起来, 转化为实际电子产品的综合设计能力, 这对于培养电子类大学生的创新能力具有非常重要的作用。

1 实验内容的设计

对于电子设计综合实验这门课程而言, 是有别于其他电子类实验课程的, 也不同于传统

的基础性和验证实验，它更加注重对学生创新思维、创新精神和创新能力的培养，鼓励学生充分发挥自己的想象力与创造力，在实验中培养个性，并使其养成良好的实验习惯。作为全院公共选修课，由于课时有限，我院电子设计综合实验为 31 学时，教师讲授所占学时较少，只是在实验过程中，根据出现的问题适时进行讲解，理论授课时间一般不会超过 10 个学时，综合实习不超过 3 学时。课程的大部分时间留给了学生，教师在实验过程中只起引导作用。

我们编写了实验指导书，将实验内容分为五个模块，即：电源模块的设计、输入通道的设计、输出通道的设计、显示单元的设计、综合系统的设计与调试。这五个模块把电子类学生所学的电路分析、模拟电子技术、数字电子技术、单片机原理与应用、接口技术、Protel、电子电路仿真技术等专业知识进行进一步的优化组合，从应用的角度去学习、去理解、去使用。在模块的制作过程中，学生们还要用到台钻、热熔胶枪等工具，实验过程更加实战化，实验现场更加接近工厂环境。在实验内容上，集电路设计、电路版图制作、元器件焊接、系统调试为一体，既包含了对理论知识的理解与掌握，又增强了学生的动手能力，为大学生创新性思维培养与创造能力的发挥提供了一个实践平台。电子设计综合实验教学内容及安排如表 1 所示。

表 1 电子设计综合实验教学内容及安排

教学主题	课时	教学内容	作业
理论教学	10	掌握电子元器件、万用表、示波器、台钻、电烙铁等的使用方法	查阅资料
综合实习	3	焊接实习、钻床使用	
电源模块的设计	3	设计集成直流稳压电源	电路设计与焊接调试
输入通道的设计	3	了解单片机输入通道的工作原理，掌握常用单片机输入通道的电路设计方法	转速测量 键盘输入 电容测量
输出通道的设计	3	了解单片机输出通道的工作原理，掌握常用输出通道的电路设计方法，包括功率放大电路的原理和设计调试方法，电机控制电路的设计、组装与调试方法	设计直流电机控制电路 设计步进电机控制电路
显示单元的设计	3	熟悉液晶显示原理，了解显示接口工作原理，掌握 LED 数码管显示方法及其编程	设计七段字型数码管器件、 液晶显示器件驱动程序
综合系统的设计及调试	6	全面理解并掌握基于单片机的电子系统的设计方法、原理、流程、电路结构及调试方法	将以上各部分组成一个完整的控制系统，并进行联调

2 实验功能的实现

在实验的功能实现上，鼓励学生采用模块化设计思想，利用总线扩展的方式连接集成。单片机最小系统需要学生自己动手焊接制作，若采用 DSP、FPGA 等芯片，可使用成品最小系统板，其他功能部分，如 A/D、D/A、电机驱动、显示电路等均需要设计独立模块，在 Protel 中绘制电路图，在万用板上实际焊接电路。学生可以在课程开始前自由结组报名，每组有同学 2—3 人，题目可以是表 1 中几个单元模块的叠加，也可以根据组内同学的情况自行设计实验题目，并由指导教师审阅批准。如，在实验中，有一组同学设计了一个恒温容器控制闭环

系统,要求温度始终保持在 25°C ,他们用纸盒做一个小的容器,内放电灯加热,风扇通风降温,用热敏电阻测量温度变化。学生在实验中先查阅相关文献资料,再在了解基本设计框架的基础上采用小组内讨论方式,形成一个总体方案,并进一步研究技术细节,包括电机功率选择、功放电路设计、传感变送设计、显示方案等。教师充当评判与指导的角色,不参与方案的设计,只是对实验的方案是否可行,以及学生在实验过程中出现的不良实验习惯与实验方法进行纠正。在电子设计综合实验教学中应遵循先易后难,理论与实践相结合的原则,充分发挥大学生的主观能动性,变被动的重复性实验为主动的创新性研究,让学生多思考、多动手。对于学生在实验中遇到的各种问题,教师不是直接给出答案或结果,而是采用引导的方式,让学生去查资料,找原因,克服学生的畏难情绪,提高学习的兴趣,在学生找到解决方案后及时向大家公布问题的解决方案,激发学生学习的积极性,培养学生的探索精神与创新思维。

3 实验成绩的考核

电子设计综合实验的实践性非常强,与传统的理论课程和实验课程有较大的差异,学生成绩的评定方式也需要改进。传统的成绩考核为理论考试与平时成绩的综合,并不适合本门课程,因此,我们把成绩考核改为教师评分与学生互评相结合的方式,并且,实行阶段性评分制度,对阶段性的成果作一个总结,进一步体现实验成绩评定的公平公正与开放性。被考核的学生需要拿出自己的作品进行演示和介绍,教师和学生评委对被考核的学生进行提问,并对作品的完成情况、制作工艺、功能、性能指标进行综合评判,打出分数。在这种情况下,学生的迟到、早退以及旷课现象消失了,取而代之的是实验课上积极学习,课下争相讨论,课上完不成的任务课下进入开放实验室继续做,学生都进入了实验室,保证了课程的学习效果,使学生从内心深处认识到实践环节的重要性,认真做好每一阶段的实验。

4 实验室的建设

电子设计综合实验课程制定了实验目标、实验内容、实验要求等项目,但没有明确实验的具体方法、步骤,这就要求学生自己去规划实验思路、设计实验方案。在实验开设之初,存在学生不知所措的情况,在完不成实验内容的情况下,开放实验室成了学生们的首选。开放实验室应结合当前实验室所开出的课程以及大学生科研情况,调配相关仪器设备,组织教师进行指导。在开放实验室建设中,鼓励教师进行教改立项,如教师带领学生科研攻关小组开发了学生实验刷卡登记系统、实验仪器、工具、元器件管理系统,引入了计算机网络管理模式,方便学生实验设备借用,学生可以根据实验室、教师值班情况进行实验预约,在预约的时间内进行实验,不仅在形式上开放了实验室,并且在时间上、空间上、设备上、管理上真正做到了对学生的全面开放。

5 实验效果

从 2007 年开始,电子系开始将电子设计综合实验与中国电子学会电子设计工程师认证相

结合，将实验内容重新整合，采取课内学习，课外认证的作法，取得了较好的成绩。从考试情况来看，电子工程师考试通过率由 2008 年秋季的 89.8%，提高到了 2009 年春季的 100%，通过率有明显提升，但从理论成绩与实操成绩平均值来看，分数却是呈下降趋势，说明考试题目难度在逐渐增大，这一点也能从获得助理认证资格的人数可以看出；从理论成绩标准差和实操成绩标准差可以看出，从 2008 年秋季考试到 2009 年春季考试，学生的整体水平分散性逐步缩小，说明经过课程的整改，学生的总体水平呈学步上升趋势，如表 2 所示。

表 2 电子设计工程师认证考试成绩

考试时间	主考年级	报名人数	缺考人数	通过率	理论平均/标准差	实操平均/标准差	助理人数
2008 年秋季	2005 级	59	0	89.8%	71.49/5.77	73.22/13.04	2
2009 年春季	2006 级	40	1	100%	64.37/3.14	71.22/3.19	0
2009 年秋季	2007 级	69	0	97.1%	61.25/3.57	75.89/10.99	0

6 结束语

电子设计综合实验教学改革是结合当前我院实际情况进行的初步尝试。从教学效果来看，这种开放式的教学方式有利于提高学生的综合素质，激发学生的创新意识，培养学生的创新能力。将部分设计性实践类课程与开放实验室相结合，促进了实验室的建设，增强了实验教师教改的信心，为电子类综合实验教学改革的向前推进提供了动力支持。

参考文献

[1] 张学军. 电子综合实验系统设计与实现[J]. 实验技术与管理, 2005, 22(5): 55—58.

[2] 王书纯, 杨艺华. 电工电子实验教学提高学生动手能力初探[J]. 实验技术与管理, 2005, 22(11): 119—120.

[3] 王海梅, 全卫强, 王兴君. 我院电子信息工程技术专业的探索与改革[J]. 陕西国防工业职业技术学院学报, 2008, 18(2): 35—37.

[4] 翟红云, 凌艺春. 电子设计竞赛促进下的单片机教学改革初步探索与研究[J]. 广西大学学报(自然科学版), 2008, 33(S): 365—367.

[5] 吴水根, 龚建萍, 江延湖等. 电子设计与制作课程的教学研究[J]. 江西教育学院学报(综合), 2008, 29(6): 19—21.

[6] 李莉, 路而红. 电子信息工程专业学生创新能力的培养[J]. 北华航天工业学院学报, 2008, 18(S): 59—60.

作者简介

王建新（1977—），男，河北人，工学博士。北京电子科技学院电子信息工程系讲师。主要研究方向：检测技术、信号处理。

研究性教学模式下电子信息类课程教学改革探索

王建新 李秀滢 周玉坤 陈汉林

(北京电子科技学院, 北京, 100070)

摘要: 研究性教学是教学改革中出现的一种新的教学理念和教学模式, 强调过程、应用、体验, 有利于激发学生的学习兴趣, 培养学生的学习能力、研究能力和创新能力。本文从研究性教学的特点及模式出发, 结合电子信息类课程的自身特点, 探讨了在电子信息类课程中开展研究性教学的方法, 为构建合理的电子信息类课程教学体系提供借鉴和帮助。

关键词: 研究性教学; 电子信息; 教学改革

Study on Electronic and Information Courses Research-based Teaching Mode

WANG Jian-xin LI Xiu-ying ZHOU Yu-kun CHEN Han-lin

(Beijing Electronic Science and Technology Institute, Beijing 100070, China)

Abstract: Research-based teaching is a new teaching method and teaching mode in the teaching reform. Research-based teaching emphasizes the process, application and experience, which is advantageous to stimulate the students' study interest, to improve the learning, research and innovation ability. This paper studies the characteristics and mode of research-based teaching, discusses the teaching method of the electronic and information courses, in order to construct a reasonable teaching system.

Keywords: research-based teaching, electronic and information, teaching reform

长期以来, 由于受到教学条件、师资力量等限制, 教师忙于完成教学计划与教学任务, 开展研究性教学一直限于理论研究与小范围尝试状态, 并且教学内容呆板、模式单一, 实验繁杂, 但普遍侧重于基础性验证实验, 综合性、设计性实验内容少, 只重视学生实验结果, 忽视教学过程。这种传统的教学模式限制了学生的个性发展, 制约了学生创新精神和创造力的培养, 教学效果较差, 多数学生对学习的积极性不高、兴趣不浓。为了改变传统的教学模式, 提高学生综合素质及创新能力, 培养适应社会发展的高质量人才, 应探索一种师生互动、学生主动的研究性教学模式。

1 电子信息类课程研究性教学的特点

研究性教学又称为主题研究、项目课程，是在教师的指导下，学生从学习生活和社会生活中选择并确定研究专题，用类似科学研究的方式，主动地获取知识，应用知识，解决问题的教学模式。它是应对新世纪知识经济的挑战而发起和实施的一种新的教学模式，对于激发学生的学习兴趣，培养学生的创新意识与能力具有积极的作用。

与一般教学中只重视学生学习的量化结果不同，研究性教学更加重视课程教学的过程，把重心放在学生的学习过程中，注重引导与改善学生的学习习惯和提高认知的思维水平。此外，对于教学效果的评价也不仅仅停留在考试成绩的层面，而应着重考察学生在利用现代信息技术手段和科学研究的基础上进行调研、实验、分析的方法，从而进行自主判断、选择、解释和应用，达到发现问题、分析问题和解决问题的目的。在教学中要做到学生积极参与，分工协作，在实验过程中应注重应用，做到学以致用，重视在学习过程中对原理和现象的认知过程，把感性认识提升到理性认识，了解基本原理，掌握基本方法，运用基本手段，让学生亲身参与实践活动，在体验、内化的基础上，逐步形成自觉指导学习行为的个人观念体系。

2 电子信息类课程研究性教学的模式探索

2.1 课堂教学

在课堂教学中，应用研究性教学理论，引导学生在教师讲授的过程中思考，鼓励学生课堂提出问题或提出大胆设想。如，在《信号与系统》教学中，可以提出：小信号可以放大，频率能被放大吗？信号在放大后，频谱会发生变化吗？学生一般会先凭直觉经验得到一个答案，再找到公式进行分析，得出一个结论。如果教师直接将结论或答案给出，然后再进行分析不利于学生主观能动性的发挥，起不到激励学生的问题意识，因此，可以让同学们相互讨论几分钟，再让学生们主动回答问题，充分保护与诱导学生的探索意识，满足学生的求知欲望。最后，可以利用信号发生器与频谱分析仪进行结论验证，从实践的角度加深学生对于信号的认识，激发学习兴趣、问题意识，培养探索与创新精神。

2.2 实验教学

实验是电子信息类课程教学环节中非常重要的一环。传统的电子信息类实验教学项目多为理论课程问题的验证，让学生掌握基本的分析方法，其实验内容受到极大限制，学生在规定时间内完成规定的实验步骤，内容老化，形式僵化，学生处于被动地位，不能充分发挥学生的积极性与创新性，并且考核成绩经常与完成的时间挂钩，使学生即使发现了一些问题也无心深入研究，对学生的问题意识、探索与创新精神培养十分不利。应提高综合性、创新性实验的比例，鼓励在实验课上发现并讨论问题，对于课堂内没有做完的实验允许课下继续完成，对提出问题的学生给予肯定，对改进实验方案学生给予表扬，并在成绩上有所体现。

2.3 开放实验

进行开放实验室建设，组织实施开放性实验。对于勇于探索、学有余力的学生，鼓励进

入开放性实验室。开放实验室可以设置一些开放实验题目,并且允许学生自主命题,对自己感兴趣的问题进行实验研究。学生可以自由安排时间和内容进行实验,实验内容由实验室负责教师审察,应体现设计性、综合性和工程应用性。在实验之前,应让学生拟定实验的具体做法与步骤,分析可能出现的现象,给出预想结果。教师应对学生的方案进行综合评估,指出实验中的错误做法和不良实验习惯,防止仪器仪表损坏,降低实验结果的不确定性,并与学生共同探讨问题解决方案,培养学生的团队意识、动手能力、创新能力。在实验成功之余,对实验的过程与结果进行总结,得出可信的结论,使学生获得一种成功的满足感,有利于实验兴趣的激发。

2.4 综合实训

电子信息专业培养具有电子电路与信号分析专业知识,具备从事电子系统设计、数字信号处理和技术服务工作的高素质高技能人才。综合实训应改革统一模式的教学方式,学生可根据自身的兴趣爱好与能力水平来选择如何完成实训项目以及怎样完成实训项目,应体现区别对待,差别培养的原则。动手能力较强的同学在完成实训项目后可对实验进行再设计并添加相应的新功能,动手能力较弱的同学可采取先观摩他人实验,在掌握原理、明确方案的基础上再进行实际操作。实训项目分基础部分与提高部分,使学生有更多自由发挥的空间,强调对学生问题的认识、分析、思考能力的培养,为创新性思维的发展打下良好基础。

3 电子信息类课程研究性教学的措施

3.1 实施方案

在教学过程中,使学生形成主体意识,积极地投入到教学活动中,避免传统教学中只重教师讲授的缺点,形成学生学习责任感,降低对老师的依赖心理。明确研究性教学的目的、意义、任务、要求、方案,使学生对研究性教学有一个正确认识。

研究性教学本着发挥学生的特长为原则,充分发掘学生自身潜力,以组为单位,分工合作。教师针对教材、实验等为学生提供一些思考题,并鼓励学生在此基础上进行深入讨论与研究,查阅相关文献资料,找出问题的关键,提出解决方案,并对一些可实验操作的问题进行实验验证。在资料搜集与整理过程中,鼓励学生使用各种电子文献检索方式,如:CNKI、维普资讯、万方数据等。对于组内的每个成员都应给出对问题的分析报告,在小组内进行讨论,形成初步方案,再进行深入讨论,形成结论。研究性教学注重对学生参与学习的全过程评价,并考虑学生的个体差异。将小组的评价与个体评价相结合,强调对小组的评价,并在此基础上根据个人的表现给出成绩。由于学生的学习经验、水平、理解能力等的限制,对某些问题的认识可能出现疏忽和理解偏差等情况,教师必须进行建议性的纠正、补充、归纳等工作,帮助学生形成明确、系统、深刻的认识,从而形成合理的学习观念。教师在整个过程中扮演指导者和辅助者的角色,帮助学生解决研究过程中遇到的问题,协调各种关系,以使学生的研究活动顺利进行。

3.2 师资建设

研究性教学需要有高素质的教学团队,教师应从知识型向能力型转变,在思想上也要做

到转变,转变传统教学中注重“教”而轻视“学”的思想。教师应深入生产建设一线,及时了解行业发展动态,将最新的科学技术和学术理念带入课堂,渗透到教学中,这样既能培养学生的兴趣,又能提高学生的专业实践能力。在此过程中,教师要身先士卒,带头学习和研究先进的教学方法、教学理念,掌握科学的教学手段,与学生共同参与学习活动,融入学生、了解学生。

3.3 实验室建设

对于电子信息类研究性教学而言,培养学生的精神、创新能力,离不开实验室、机房、实习基地。学校应完善实验室管理机制,将实验场所进行部分或完全开放,使学生有机会进入实验室。在实验设备的配备上,学校应下大力气投入,使得仪器设备在种类、质量、数量上都能够得到保证。在时间的安排上,学生应可以预约实验时间、自定实验内容,既可以把实验课上没有做完的实验完成,又可进行探索性的科研实验。由于实验室的开放会加速实验设备老化,损坏的数量也会增加,因此,应加强设备维护队伍的配给,在情况允许下,可由教师组织学生中能力较强的同学进行设备维护、更新培训,使其掌握基本维护技能,并参与日常实验室管理与实验设备维护等。

4 结束语

研究性教学模式打破了以往以教师为主的教学模式,优化了学习过程,激发了学生学习相关课程的兴趣,在学习中相互促进、相互提高,开拓学生的创新意识,创设研究情境,使学生用研究性思维进行问题分析,创新教学模式。在研究性教学模式下,学生有了一个宽松、民主的氛围,使得学生的学习活动能自觉而合理的开展。应用研究性教学理念,从实际出发,采取一定的方法和措施,不断探索,切实以提高学生的综合素质与应用能力为出发点,把能力培养落到实处。

参考文献

- [1] 刘运智. 论高校研究性教学与研究性学习的关系[J]. 中国大学教学, 2006, (2): 24—34.
- [2] 马东堂, 王杉. 通信基础理论课的研究性教学探讨[J]. 高教论坛, 2008, (2): 66—68.
- [3] 王红玲. “研究性”教学模式的探索与应用[J]. 科技咨询导报, 2007, (14): 248.
- [4] 王海梅, 全卫强, 王兴君. 我院电子信息工程技术专业的探索与改革[J]. 陕西国防工业职业技术学院学报, 2008, 18(2): 35—37.
- [5] 黄亚平. 电子信息专业教改的探索和实践[J]. 广东白云学院学报, 2008, 15(2): 55—58.
- [6] 黎小桃. 应用研究性教学理念培养学生的创新能力[J]. 中国科技信息, 2005, (16): 223.

作者简介

王建新(1977—),男,河北人,工学博士。北京电子科技学院电子信息工程系讲师。主要研究方向:检测技术、信号处理。

动画自动生成中基于分镜头的摄像机规划

王巍峰

(北京工业大学计算机学院,北京市多媒体与智能软件重点实验室, 北京, 100124)

摘要: 本文以中科院陆汝钤院士提出的全过程计算机辅助动画自动生成课题为背景, 设计了一种基于分镜头的摄像机规划系统。系统以情节规划作为输入, 结合电影界的摄像知识, 分层处理情节信息。它利用 Prolog 语言构造摄像规则库, 将动画中的情节脚本转换成分镜头脚本, 为后期的摄像机定量计算提供方便。

关键词: 动画自动生成; 摄像机规划; 分镜头脚本

Camera Planning Based On Storyboarding In Automatic Animation Generation

WANG Wei-feng

Abstract: In this paper, I designed a camera planning system based on storyboarding with the background of academician Ruqian Lu's Full-life cycle computer aided animation generation subject. The system accepts plot planning as input firstly, then uses camera knowledge in file, finally slices plot information. It constructs camera rule base by using Prolog language, converts plot script in animation to storyboarding. This will be convenient for the further quantitative camera planning.

Keywords: Automatic Animation Generation; Camera Planning; Storyboarding

1 引言

随着三维动画技术被应用到更多的领域, 伴随其出现的最直接问题——摄像机规划问题也自然受到了学者们广泛的重视。摄像机规划就是解决在三维环境中, 观众的观点应该在哪里, 摄像机应该朝向哪个方向以及如何运动等问题。在一部三维动画中, 摄像机规划非常重要。因为规划结果的好坏直接影响到观众对整部动画的欣赏。一个出色的摄像机规划, 可以将动画的情节完整清晰地传达给观众, 使观众获得一场赏心悦目的视觉盛宴。

2 研究现状

摄像机有很多种规划方法, 最早的是基于图形描述的摄像机规划。其代表人物是 J Blinn^[1]

和 Drucker^[2]。Drucker 的系统是对 Blinn 的归纳, Blinn 的解决方案只能处理摄像机规划的特殊问题, 而 Drucker 通过约束和最优化操作使系统具备了摄像机初始化、摄像机控制和选择最优方案的功能。

从系统解决规划问题的机制入手, 现有系统可以划分为代数系统、交互系统。代数系统将摄像机规划问题转换成一个代数问题进行求解, 求出的结果一般会很准确, 但是缺乏灵活性; 交互系统依赖用户输入, 将其映射到摄像机的对应属性上以驱动摄像机状态的改变, 此类系统主要适用于实时三维交互系统, 如 William Bares^[3]的 ConstraintCam 系统, 其优势是操作简单, 反应迅速; 但它规划出的结果缺乏构图的观念, 不能形成连续的镜头。Tsai-Yen Li^[4]等人设计的摄像机模型也是交互式的, 它可以模仿导演、摄影师和编剧的工作。

基于摄像机属性的摄像机规划也是一种常见的方式。Languénou^[5]和 Li-wei He^[6]的系统就是这种类型。Languénou 的系统在静止时, 可以通过目标图像定位在屏幕中的位置找出摄像机的拍摄位置。在运动时, 可以将摄像机的运动定义成一个运动轨迹, 利用系统找出这个轨迹上的关键点从而确定摄像机的位置。而 Li-wei He 提出的则是一种智能实时的摄像机控制, 它遵循摄像的基本原理, 可以较好的处理拍摄手法和镜头过渡等问题。

此外, He and Salesin^[7]于 1996 年开发了一个摄像机规划系统也是这种类型的代表, 这个系统使用说明性摄像机控制语言(DCCL), 系统利用电影学的知识可以将单一镜头组合成复杂镜头, 从而实现较为复杂的摄像机规划。

3 摄像机模块概述

在陆汝钫院士提出的全过程计算机辅助动画自动生成系统中, 摄像机规划模块被定义为在摄像专业知识库的协助下, 自动对上层传入的故事背景及动作序列进行规划, 规划得出对应的摄像机动作语句序列, 通过定量计算, 将语句转化为摄像机具体的位置参数, 并在最终的动画生成中使用。

按照陆汝钫院士在《Automatic generation of computer animation》^[8]一书中关于摄像机规划的设计思想, 可以将摄像机规划模块划分为五个层次。它们分别是摄像规划要求原语(DPRS), 群组摄像原语(RCS), 高级摄像原语(HLCP), 基本摄像原语(BCP)和量化的是摄像语句(QCS)。由此可以看出, 摄像规划是一个复杂的过程, 我们可以将它的规划顺序与软件工程进行对比。

摄像规划要求原语→软件需求说明

群组摄像原语→软件系统设计

高级摄像原语→软件模块设计

基本摄像原语→高级语言编程

量化的摄像 语句→汇编语言编程

根据这种设计思想, 我将摄像机规划模块成分镜头规划和摄像机定量规划两部分。在动画自动生成系统中, 分镜头规划处在情节规划之后, 它接收情节规划的信息, 结合摄像的知识, 将情节按照一定的规则划分成分镜头脚本, 传送给下一层。它属于定性的内容。此后, 为了使自动生成的动画能够最终展现出来, 要对定性分镜头进行定量规划。在进行摄像机定量规划时, 需要事先知道具体的背景、人物和物体的位置作为参考坐标系, 只有它们的定量位置确定以后, 摄像机的位置才能够计算出来, 所以摄像机定量规划处在动作规划之后, 主要接收动作规划的定量信息, 然后根据这些信息, 将分镜头脚本中的摄像机位置计算出来。

4 分镜头规划

4.1 分镜头规划的设计

4.1.1 摄像规划要求原语

系统接收上层的情节规划信息以后，提取情节类型、情节所涉及的角色信息等内容，根据归纳出的导演知识，产生摄像规划要求原语。它将作为摄像机规划模块的输入。

摄像规划要求原语主要包括角色数量、角色姓名、角色地位、角色朝向、角色位置和情节内容等几部分信息。其中，角色地位指的是人物之间的关系，同一个画面中的几个人物，有可能人物之间的重要性是平等的，如两人面对面的对话，也有可能人物之间的关系有主次之分，如演讲中的演讲者与听众；角色位置此时给出的是定性的信息，主要是标明人物相互之间的关系，如两人谈话时的面对面，两人追逐时的一前一后等等；情节内容是规划的重要依据之一，分镜头规划会根据不同的情节内容，产生不同的分镜头方式，主要的情节包括谈话、追逐、会议等等。除了上述这些必要信息外，摄像规划要求原语还可以根据用户的输入对节奏和摄像机视角进行设定，不同的节奏和摄像机视角，可以产生不同的分镜头规划结果。

摄像规划要求原语的规则主要按照角色的数量进行划分。

比如，情节是角色 A 在跑步。进行处理后产生的摄像机要求原语为：

角色数量：1
角色姓名：A
角色地位：平等
角色朝向：前
情节内容：跑
角色位置：空
节奏：快/正常/慢
摄像机视角：近景/远景

再如，情节是角色 A 和角色 B 是在面对面的谈话。进行处理后产生的摄像机要求原语为：

角色数量：2
角色姓名：A
角色地位：平等
角色朝向：朝向 B
情节内容：谈话
角色位置：面对面
角色姓名：B
角色地位：平等
角色朝向：朝向 A
情节内容：谈话
角色位置：面对面
节奏：快/正常/慢
摄像机视角：近景/远景

4.1.2 群组摄像原语

系统根据摄像规划要求原语内容，按照规则将其翻译成群组摄像原语，群组摄像原语比较简单，它主要标明了用户可以在场景中看到哪些内容，它并不包含任何摄像的技术。

如上面的两个例子，角色 A 奔跑可以转化为 `running(A)`、`running (A,quick)`等情况，分别表示普通状况下 A 在奔跑和快节奏下 A 在奔跑；角色 A 和角色 B 面对面谈话可以转化为 `face_to_face (A,B,MCU)`、`face_to_face (A,B, VLS)`等情况，分别表示近景拍摄 A 和 B 的谈话以及远景拍摄 A 和 B 的谈话。

4.1.3 高级摄像原语

高级摄像原语是拍摄动画的一个基本单位。它将群组摄像原语中所描述的内容，用摄像技术加以实现，它是由导演和摄像师的实践经验总结出来的。分镜头系统规划到这里已经开始涉及具体拍摄技巧，在高级摄像原语中，包含了镜头景别、镜头角度等分镜头脚本中所必不可少的信息，但是它所包含的信息并不完全满足分镜头脚本的需要。

群组摄像原语向高级摄像原语的转化是根据 Prolog 规则进行处理，下面我将以 `running(A)` 和 `face_to_face (A,B,MCU)`为例，表明高级摄像原语的输出结果。

`running(A)`

镜头号：1
镜头名称：侧面拍
拍摄目标：A
拍摄景别：近景
拍摄角度：水平
时间：long

单人情节较为简单，因此一般以一个镜头为主。

`face_to_face (A,B,MCU)`

镜头号：1
镜头名称：外反拍
拍摄目标：A
拍摄景别：近景
拍摄角度：水平
时间：a little long
镜头号：2
镜头名称：外反拍
拍摄目标：B
拍摄景别：近景
拍摄角度：水平
时间：a little long

两人面对面对话时可以有多种拍摄方式，如图 1 所示。目前暂选最常用的一种方式进行示例。

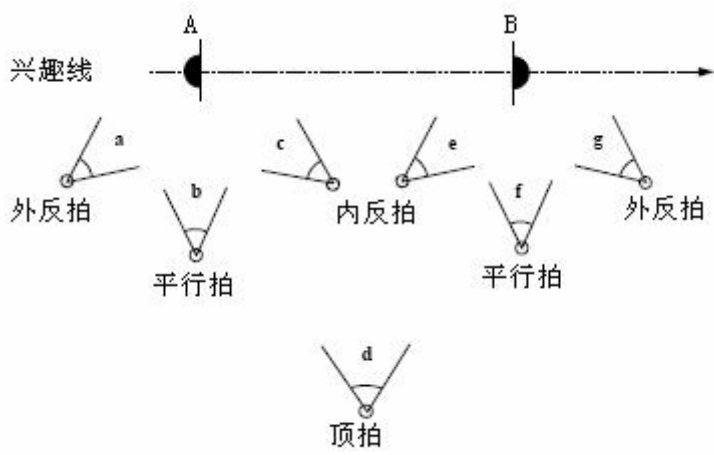


图 1 面对面对话时常用的拍摄机位^[7]

4.1.4 基本摄像原语

为了使分镜头脚本更加完整，同时也是为使计算机能够更好地处理规划出的分镜头信息，分镜头规划要进行第四步操作，即从高级摄像原语向基本摄像原语转化。基本摄像原语是将高级摄像原语进一步细化，每个基本摄影原语都包含摄像机的一次动作，如平移、推进、拉出、摇摄和固定镜头等。一般，基本摄像原语的参数包括被摄目标、景别、方向、俯仰角度、摇摄角度、速度参数、持续时间等内容。通过这部分的规划，我们最终可以获得从情节规划到分镜头规划的规划结果，得到一个完整的分镜头脚本。

与前几次转化一样，这步转化也是根据 Prolog 规则进行处理，最终完成定性部分的输出。同样以 running(A)和 face_to_face (A,B,MCU)为例：

running(A)
镜头号：1
镜头名称：侧面拍
拍摄手法：跟
拍摄目标：A
起始景别：近景
终止景别：近景
拍摄角度：水平
开始时间：t1
结束时间：t5

face to face (A,B,MCU)

镜头号: 1
镜头名称: 外反拍
拍摄手法: 定
拍摄目标: A
起始景别: 近景
终止景别: 近景
拍摄角度: 水平
开始时间: t1
结束时间: t3
镜头号: 2
镜头名称: 外反拍
拍摄手法: 定
拍摄目标: B
起始景别: 近景
终止景别: 近景
拍摄角度: 水平
开始时间: t3
结束时间: t5

4.2 分镜头规划的实现

前文中已经提到，分镜头规划中各原语之间的转换是依靠规则库来完成的。本系统中的规则库是按照 Prolog 规则进行处理的。Prolog 作为一种逻辑编程语言，它建立在逻辑学的理论基础之上，广泛地应用在人工智能的研究中，可以用来建造专家系统、智能知识库。Prolog 相关的编译器有很多，本系统选取了 SWI-Prolog 作为 Prolog 规则处理的编译器，它具有运行速度快，可移植性强等特点，同时它也为 Java 等语言提供了良好的接口。

分镜头规划的重点在于规则的设计与书写，以两人面对面对话这个情节为例，通过情节规划，我们得到群组摄像原语 face_to_face (A,B,MCU)。之后我们会通过规则库查找可以与这个群组摄像原语相对应的规则，并按照这个规则产生对应的高级摄像原语。在规则库中，书写了如下的规则：

storybording([shot_id:1,shot_name:外反拍,shot_target:X,shot_view:近景, shot_direction:_,shot_angle:水平,shot_time:a_little_long],[shot_id:2,shot_name:外反拍, shot_target:Y,shot_view:近景,shot_direction:_,shot_angle:水平,shot_time:a_little_long]) :-face_to_face(目标 1,目标 2,MCU).

即当出现形如 face_to_face(目标 1,目标 2,MCU)这样的群组摄像原语时，可以产生这样一个初始分镜头，它的拍摄手法、拍摄景别、拍摄角度、拍摄时间等内容都可以得到确定。

所以当系统得到 face_to_face (A,B,MCU)时，可以产生如下的结果(因为暂时未设定镜头的朝向，所以默认为空)：

```
镜头 1 = [shot_id:1, shot_name:'外反拍', shot_target:A, shot_view:'近景',
  shot_direction:_G404, shot_angle:'水平', shot_time:a_little_long],
镜头 2= [shot_id:2, shot_name:'外反拍', shot_target:B, shot_view:'近景',
  shot_direction:_G446, shot_angle:'水平', shot_time:a_little_long].
```

这就是高级摄像原语的内容。同样的，利用 Prolog 规则可以将高级摄像原语转化成基本摄像原语，转换规则书写如下所示：

```
output([shot_id:1,shot_name:外反拍,shot_methods:定,shot_target:目标 1,shot_view_start:
近景,shot_view_end:近景,shot_direction:_,shot_angle:水平,shot_time_start:t1,
shot_time_end:t3],[shot_id:2,shot_name:外反拍,shot_methods:定,shot_target:目标 2,
shot_view_start:近景,shot_view_end:近景,shot_direction:_,shot_angle:水平,
shot_time_start:t3,shot_time_end:t5])
:-storybording([shot_id:1,shot_name:外反拍,shot_target:目标 1,shot_view:近景,
shot_direction:_,shot_angle:水平,shot_time:a_little_long],[shot_id:2,shot_name:外反拍,
shot_target:目标 2,shot_view:近景,shot_direction:_,shot_angle:水平,
shot_time:a_little_long]).
```

系统通过调用这个规则，产生的输出：

```
镜头 1 = [shot_id:1,shot_name:外反拍,shot_methods:定,shot_target:A,shot_view_start:近
景,shot_view_end:近景,shot_direction:_G404,shot_angle:水平,shot_time_start:t1,
shot_time_end:t3],
镜头 2= [shot_id:2,shot_name:外反拍,shot_methods:定,shot_target:B ,shot_view_start:近
景,shot_view_end:近景,shot_direction:_G466,shot_angle:水平,
shot_time_start:t3,shot_time_end:t5].
```

基本摄像原语生成，分镜头规划完成。

5 总结和展望

本文按照陆汝钤院士在《Automatic generation of computer animation》一书中关于摄像机规划的设计思想，实现了从情节规划到分镜头脚本生成的规划过程。其中，摄像机分镜头规划需要借助现实生活中导演与摄像的实践经验，将它们转化成规则，并按照这些规则对接收的信息进行划分、加工，最终生成一系列对应的分镜头脚本。

为了实现这个目的，最重要的就是要建立一个包含丰富拍摄经验的导演规则库及其对应的拍摄规则库，系统利用了 Prolog 语言对这些规则进行表示，每一组规则都可以对应一系列的分镜头规划，这样可以保证规则库具有良好的可扩充性。同时，由于这种摄像机规划是分层进行、逐步转化的，因此在每一层的规划处理中，都可以将系统的时间复杂度控制在较低的范围内，从而使得整个系统的运行效率非常高，可以在短时间内得到规划结果。而且，正是基于这种分层的设计，可以让用户在每个层次中都根据自己的需要进行一定的修改，使系统更具灵活性。例如，在处理两人面对面对话时，默认选择使用近景进行拍摄，用户通过观看实际效果，感觉景别的选择不合适，希望使用远景拍摄，此时不需要重新进行规划，只需要在群组摄像原语层，将 face_to_face (A,B,MCU)修改为 face_to_face (A,B,VLS)，系统就会根

据用户修改产生新的分镜头脚本。

不过由于目前系统所包含的拍摄规则有限,因此只能使用简单的拍摄手法进行拍摄,这需要在未来的工作中继续学习摄像知识,丰富拍摄规则。此外,在后续工作中,我还要为系统增加镜头衔接、镜头避障等功能。

参考文献

- [1] Blinn, J.. Where am I? what am I looking at? [J] IEEE Computer Graphics and Applications, pages 76-81.
- [2] S. M. Drucker and D. Zeltzer. Intelligent Camera Control in a Virtual Environment. [M] In Proceedings of Graphics Interface '94, pages 190-199. Morgan Kaufmann Publishers. 1994.
- [3] William H. Bares, Somying Thainimit, Scott McDermott. A Model for Constraint-Based Camera Planning. [J] In Smart Graphics: Papers from the AAAI Spring Symposium (Stanford, March 20-22, 2000).
- [4] Tsai-Yen Li, Xiang-Yan Xiao. An Interactive Camera Planning System for Automatic Cinematographer. [J] IEEE 2005 Proceedings of the 11th International Multimedia Modelling Conference (MMM'05).
- [5] Jardillier, F. and Languénou, E. (1998). Screen-space constraints for camera movements: the virtual cameraman. [J] In Eurographics '98, volume 17, pages 174-186. Computer Graphics, Special Issue.
- [6] Li-wei He , Michael F. Cohen , David H. Salesin. The virtual cinematographer: a paradigm for automatic real-time camera control and directing. [J] Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, p.217-224, August 1996
- [7] D. B. Christianson, S. E. Anderson, L. He, D. H. Salesin, D. S. Weld, and M. F. Cohen. Declarative Camera Control for Automatic Cinematography. [J] In Proceedings of the American Association for Artificial Intelligence 1996, pages 148-155, 1996.
- [8] Ruqian Lu, Songmao Zhang, Automatic generation of computer animation. [M] LNAI 2160. Springer Verlag, to be published.

作者简介

王巍峰,男,北京人,1984年9月4日生,北京工业大学在读硕士研究生,研究领域:人工智能,摄像机规划

机场柜台资源共享解决方案

—中国民航离港多主机共用平台系统

王欣明 高 新

(中国民航信息网络股份有限公司)

摘 要: 针对机场资源日益冲突的问题, 实现了一种新的离港业务系统, 并研究了传统协议和开放 (MATIP) 协议的主机接入方法, 以及基于 NIO 的多客户端接入等关键技术。该系统无论在技术实现上还是在业务处理上均达到了国际先进水平, 现已经成功投产了多家机场航空公司, 并获得了中国民用航空局颁发的科学技术成果鉴定证书和中国民航科学技术奖二等奖。

关键词: 离港系统; 传统协议; MATIP; 多主机共用平台

引言

随着国际航空运输业、旅游业的重组与联盟趋势加剧和《中美扩展航空服务协定》、香港、澳门特区航空运输合作备忘录的签订, 越来越多的国外航空公司开始在中国国内机场开设通航点, 从而大幅度提升了往来境外的航班的数量和密度。再加上基于客源、联盟, 以及最新的市场运营模式, 航空公司的业务越来越细分, 服务也越来越细分, 个性化趋势逐步明显。

1 国际机场资源冲突催生离港新业务模式的出现

航空公司对个性化服务的推崇对为航空公司提供服务的模式提出了新的挑战, 但是作为航空公司离港系统落地载体的国际机场则面临着航空公司离港系统各不相同和航空公司个性化服务间千差万别的巨大困扰。不同的航空公司离港系统和个性化服务差异都分别向中国的机场提出了不同的资源需求。当然, 假设在不计成本和资源无限充足的情况下, 机场可以为每个航空公司提供独占使用的资源以满足航空公司需求的增长。然而在现实中, 各个机场却是要求在成本核算的可行范围内利用有限的资源解决这一困扰难题。

这就要求在机场资源一定的情况下, 出现一种全新的离港业务模式——多主机共用平台模式, 能够将各个航空公司离港主机系统和个性化服务容纳在一个系统框架下, 以达到“求同存异”的效果。针对这种新的离港业务模式, 全球民航业务发展权威组织——国际航协 IATA 制定了一项业务指导标准 (参见 IATA RP1797), 以规划这种新业务模式的发展规划和方向。

2 离港多主机共用平台系统介绍

多主机共用平台模式的出现虽然从业务上很好地解决了机场资源冲突的问题, 但是要想

从实践上真正解决机场的困境还需要民航信息企业研发出对应的离港多主机共用平台系统。离港多主机共用平台系统顾名思义是由“多主机平台”和“共用平台”两大核心部分组成，从系统运行和维护角度考虑，同时还存在一个管理模块以管理两个平台的正常运行。其中：

1) “多主机平台”作为信息传递的通信桥梁，需要实现多种航空公司离港主机系统（以下简称主机）的统一接入，并向各个主机屏蔽机场实际柜台环境的差异性，从而使航空公司可以在短时间内迅速扩张多个通航点，提供统一水准的个性化离港服务⁽¹⁾

2) “共用平台”作为各个通航航空公司离港终端应用（以下简称 TE）运行的载体，为 TE 提供访问外部资源（如设备，主机）的服务，同时向机场屏蔽航空公司主机和服务的差异性，从而使国际机场能够允许多家航空公司同时通航，提供差异化的离港服务⁽¹⁾

离港多主机共用平台系统在航空信息基础建设中扮演着越来越重要的角色，是现代化机场优选的基础设施，是服务航空公司的基本手段，更是机场走向国际化发展的必然要求。

3 中国民航离港多主机共用平台系统—AngelCue系统

为了加快机场国际化的步伐，作为中国唯一民航信息服务企业的中国民航信息网络股份有限公司（以下简称 Travelsky）经过 3 多年的潜心研究和经验积累，于 2007 年启动了中国民航运输界的一个战略性项目——自主研发具有中国特色的、遵循国际标准的中国民航离港多主机共用平台系统。该项目突破了国际上的技术封锁，开拓出了一条自主创新之路，成功研发出具有自主知识产权、自有品牌的中国民航离港多主机共用平台系统—Angel Common Use Environment（以下简称 AngelCue）。该系统无论在技术实现上还是在业务处理上均达到了国际先进水平，受到了国内外航空公司和国际机场的一致好评。

AngelCue 系统主要包括机场端管理平台 ASP（Airport Service Provider），多主机接入平台 HAS（Host Access Server）中央管理平台 CMP（Central Management Platform）和新一代航空公司离港终端应用 NG TE（Next Generation Terminal Emulator）四部分⁽²⁾，如图 1 所示。

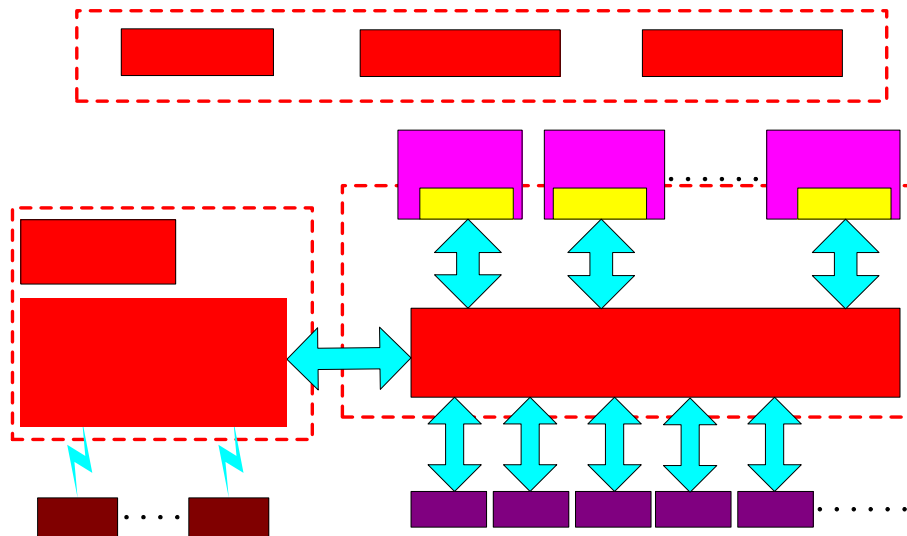


图 1 AngelCue 逻辑框架图

Angel CUE 离港多主机平台系统是一套全新开发，具有先进设计理念，符合国际民航标准的系统平台。对于机场、航空公司的主要优势集中体现在柜台资源、投资成本、系统兼容等方面。

4 AngelCue系统实现的关键技术

由于历史的原因和民航业的特点，民航主机对终端应用提供的接入协议目前处于传统协议和开放协议共存的时期。所谓传统协议就是基于 POLLING 机制的 ALC 和 UTS 协议，开放协议则是指基于 TCP/IP 网络的 MATIP 协议。AngelCue 系统解决的是 M 个航空公司离港前端应用接入 N 个航空公司主机，并且共享机场前端资源的问题。而作为一个实时性要求极高的离港系统，如何处理 M 个客户端应用和接入 N 个主机则是 AngelCue 系统实现的关键技术。

1) 基于传统协议的主机接入功能实现

民航中使用的传统协议是哑终端时期遗留下来的基于低速模拟线网络通信协议。在这种协议中，客户端只能是作为被动接受者，等待主机服务端发送的 POLL 帧，然后根据 POLL 携带的状态和客户端需要发送的信息按照协议规定的机制反馈给主机服务端⁽³⁾。如图 3 所示的 UTS 协议定义的状态转换。

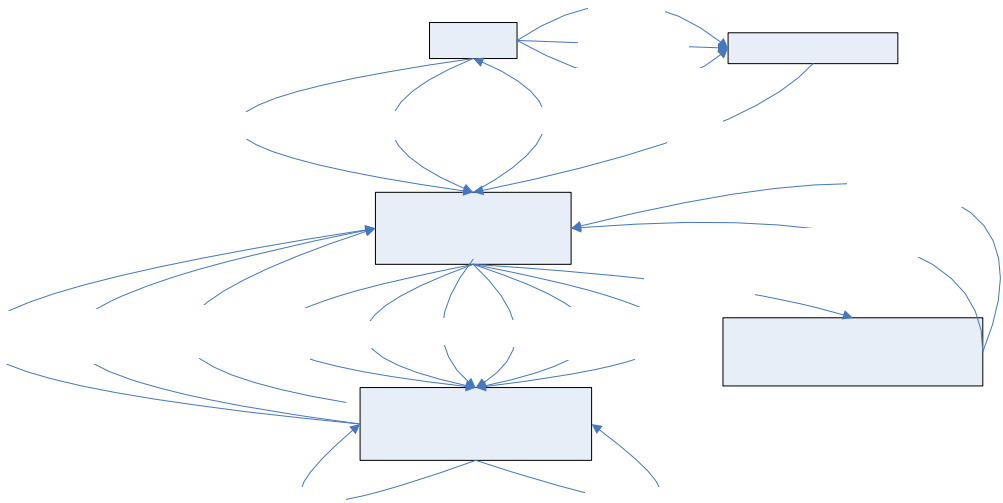


图 2 UTS 协议状态转换图

传统协议在实际应用中要依赖于一条条物理存在的低速模拟线。每条低速模拟线上都配置有不同的主机地址，相互之间是独立不干涉的。因此 AngelCue 在设计的时候选择采用进程服务组的方式，只需要开发一个传统协议接入服务实体，但根据实际使用的低速模拟线的数量来部署多个服务进程的拷贝。如图 3 所示的服务实体的逻辑结构图。

该服务实体采用基于 Java NIO 技术的 NioWorker - NioSession 服务器模型⁽⁴⁾。通过使用 NioWorker 类并实现 NioSession 接口，可以利用有限的线程来实现多客户端的处理，同时为了进一步提高处理效率，系统使用了 WorkQueue 线程池来进行具体任务的处理。

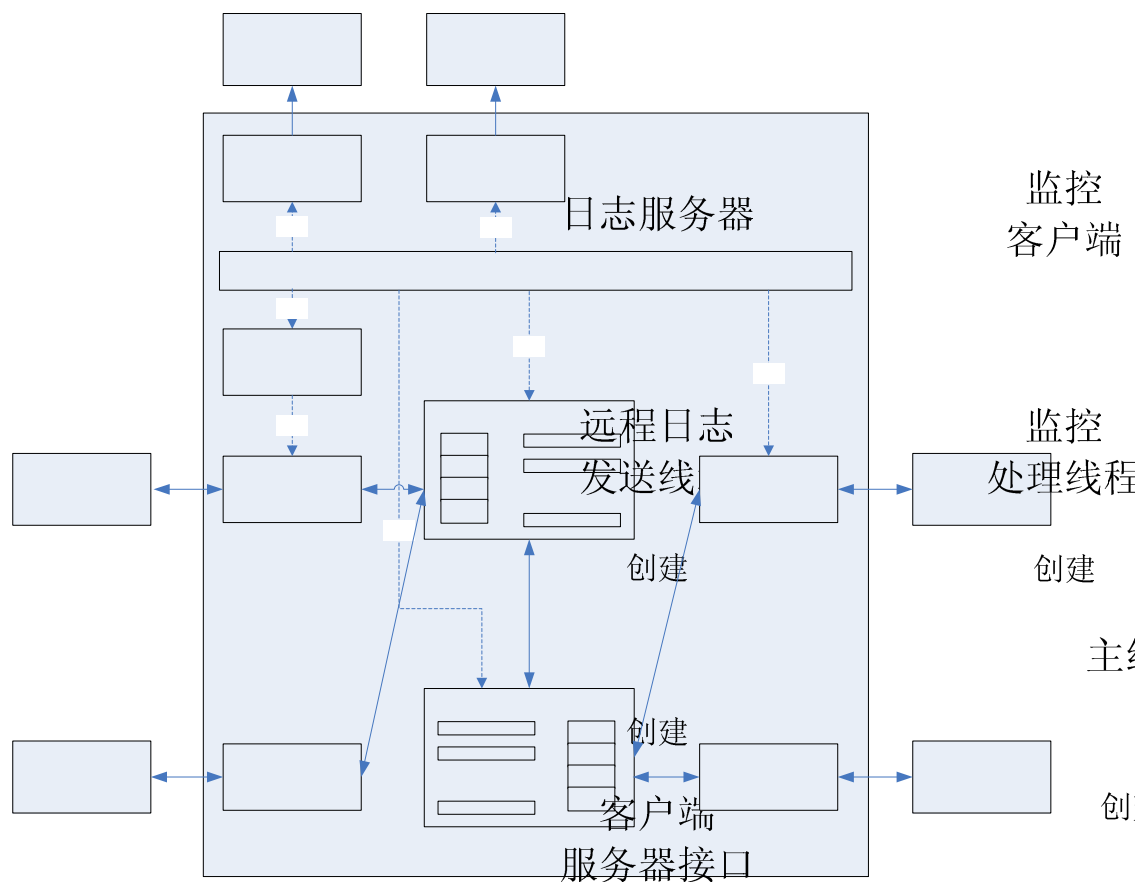


图3 传统协议接入服务逻辑结构图

主要的 NioSession 接口有客户端服务器 NioSession、客户端通讯 NioSession 和主机端通讯 NioSession。客户端服务器 NioSession 的主要任务是实现多客户端的连接处理，对于每一个客户端连接建立一个客户端通讯 NioSession，并添加到客户端 NioWorker 中；客户端通讯 NioSession 的主要任务是负责客户端数据的收发和任务 Session 提交；主机端通讯 NioSession 的主要任务是负责主机端数据的收发和任务 Session 提交。主要的 WorkQueue 任务有客户端消息 Session、主机端 ALC 消息 Session 和主机端 UTS 消息 Session。客户端消息 Session 主要负责客户端 XML 数据的解析和处理；主机端 ALC 消息 Session 主要负责主机端 ALC 协议数据的解析和处理；主机端 UTS 消息 Session 主要负责主机端 UTS 协议数据的解析和处理。主线程负责创建并初始化 NioWorker 以及 WorkQueue 线程池，同时创建独立的远程日志发送线程以及 SNMP 监控处理线程实现监控与管理。

2) 基于 MATIP 协议的主机接入功能

MATIP 是 RFC 标准文档，适用于航空通讯的标准协议。它的英文全称为“Mapping of Airline Reservation, Ticketing, and Messaging Traffic over IP”，是基于 TCP 进行数据传输的一种通讯方式^[5]。AngelCue 系统中的 MatipAdaper 模块是专门处理该协议的功能模块。线程组 MatipAdapter 的内部逻辑图。

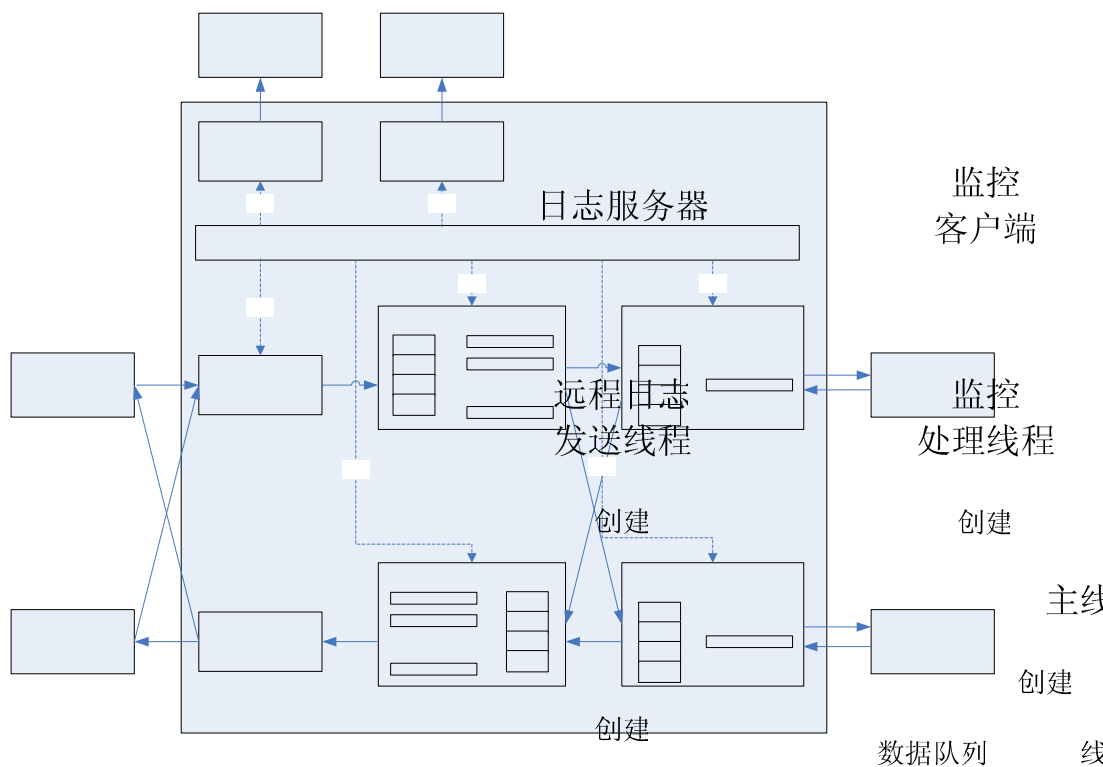


图 4 MatipAdapter 的内部逻辑图

主线程负责程序启动时的配置和初始化，依次初始化配置模块、日志模块和 MATIP 管理模块，然后进入信号处理循环。

主机端消息处理部分主要由 MATIP 连接对象模块和主机端数据处理线程池来完成，MATIP 连接对象与远程主机一一对应，单线程循环接收主机数据，并将完整的数据放入线程池的数据队列并触发条件变量，然后线程池中的线程组会对数据队列中的数据进行重组并先根据数据包头部的地址信息查找到相应客户端对象，然后由客户端对象对数据进行重组并发送给客户端。

客户端消息处理部分主要由数据接收线程和客户端数据处理线程池来完成，数据接收采用 IO 复用的单线程模型，循环接收客户端的连接和命令请求，对于新的连接，首先创建一个客户端对象实例并放入相应 MATIP 连接对象的客户端列表中，对于命令数据则会放入线程池数据队列并触发条件变量，然后线程池中的线程组会对数据队列中的数据进行重组，首先根据连接标识找到相应的 MATIP 连接对象，然后对数据进行重组并发送给主机。

远程日志处理由日志模块来完成，日志模块拥有数据队列来存贮日志信息，创建单一线程循环发送日志信息，监控处理由 SNMP 监控处理模块来完成，接受 SNMP 客户端的连接请求并创建一个接收线程来处理客户端数据，将系统信息按照 SNMP 标准发回客户端。

3) 基于 NIO 实现 SOCKET 服务器

由于业务需要，AngelCue 要实现大量客户端 TE 的接入，客户端接入层在整个系统中的地位来看，属于中心节点性质，因此该模块对前端的接入数量的处理和接入的响应速度是个关键点，特别是在有大量客户端接入的情况下。所以，这个模块的程序应该具有可用性

考虑到系统实时性的要求，AngelCue 采用 Socket 通信作为客户端 TE 接入的通信手段。在技术实现中，利用 JAVA NIO 机制，有效地实现了一个 SOCKET 服务器框架，提高了系统的接入性能和处理速度^[6]。所以，从 TE 端来看，它所连接的就是一个 Socket Server（以下称之为 ProxyServer）。图 5 是 ProxyServer 的逻辑处理流程图。

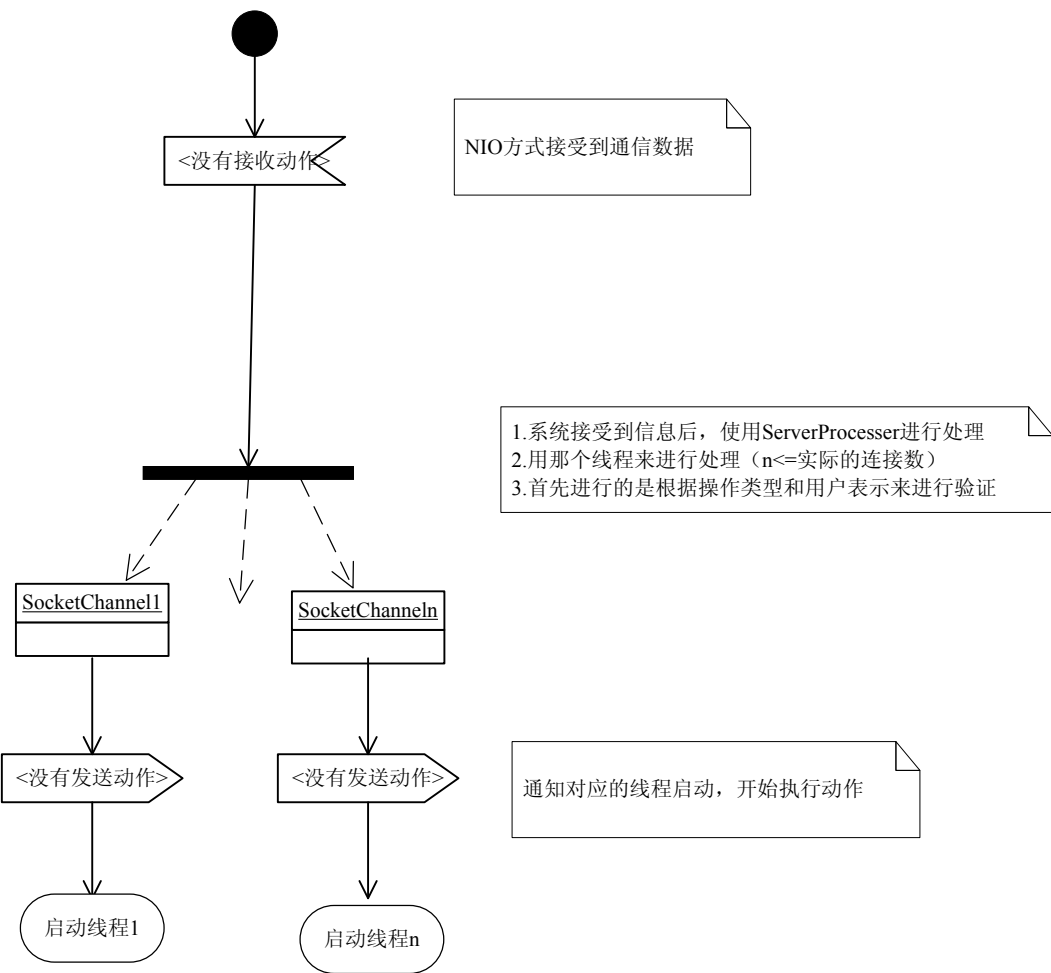


图 5 ProxyServer 的逻辑处理流程图

ProxyServer 的主要功能包括：用户认证、请求处理、请求路由、并发处理、与主机接入服务进程通信等功能，其主要功能组件包括如下：

1) ProxyServer 主要用于接收并注册 SocketChannel。对用户身份进行合法验证后，后台线程可以响应在 SocketChannel 上的消息的到来。当某个客户端对应的 SocketChannel 上有消息发送到服务端后，服务端的线程池中会有工作线程来执行对应的操作，具体的操作由 ServerProcessor 来实现。

2) ServerProcessor 的主要功能包括对消息格式的验证，消息内容的解析和后续的操作处理，然后发送处理请求到后续的主机接入服务进程中，之后工作线程会被释放到线程池中等待新的消息的到来。当后续的主机接入服务进程操作处理完成后返回结果时，会找到对应的 SocketChannel 对象，发送处理结果信息给客户端 TE。

3) 通过 JAVA 非阻塞的 NIO 处理, 配合后端线程池的方式, ProxyServer 可以同时快速处理大量客户端 TE 的并发操作。另外, 为了简化消息的处理, 进程间消息的传递采用了 XML 的封装。

5 应用案例

中国民航离港多主机共用平台 AngelCue (Common Use Environment) 系统是中国航信自主研发的达到国际领先水平的自主品牌产品。它依托于领先的 IATA 国际标准, 实现了技术盲点突破, 跻身于国际技术前沿, 填补了国内该领域的技术和产品双项空白。该产品为国内外航空公司和机场提供了完整的国际柜台共享解决方案。系统自 2008 年初投放市场以来, 已投产了多家航空公司机场, 并在提高航空公司运营效率、降低运营成本、推动机场标准化运营等方面起到了积极的作用, 形成了良好的经济效益和社会效益。目前该产品已经具备了为所有国内二线国际机场提供服务的能力, 并具备独立支持超过 40 家国外航空公司应用的能力。表 1 是截止到 2010 年 1 月, AngelCue 已经投产的机场和航空公司列表:

表 1 AngelCue 已经投产的机场和航空公司列表

<div>项目 编号</div>	机场名称	航空公司
1	天津机场 (TSN)	韩亚航空公司 (OZ)
		日本航空公司 (JL)
2	青岛机场 (TAO)	日本航空公司 (JL)
3	厦门机场 (XMN)	日本航空公司 (JL)
		菲律宾航空公司 (PR)
4	大连机场 (DLC)	韩亚航空公司 (OZ)
		全日空航空公司 (NH)
		日本航空公司 (JL)
5	烟台机场 (YNT)	韩亚航空公司 (OZ)
6	长春机场 (CGQ)	韩亚航空公司 (OZ)
7	威海机场 (WEH)	韩亚航空公司 (OZ)
8	杭州机场 (HGH)	日本航空公司 (JL)
		台湾长荣航空公司 (BR)
		台湾立荣航空公司 (B7)
		全日空航空公司 (NH)
		韩亚航空公司 (OZ)
9	乌鲁木齐 (URC)	哈萨克斯坦/阿斯塔那航空公司 (KC)
10	昆明机场(KMG)	泰国航空公司 (TG)
11	深圳机场(SZX)	台湾长荣航空公司 (BR)
		台湾立荣航空公司 (B7)

6 结束语

邓小平同志曾经说过“发展才是硬道理”。作为国内民航界首个自主品牌的离港多主机共用平台解决方案，AngelCue 极大地提升了国内民航企业在离港市场的竞争力，并为国内离港系统的国际化发展提供了坚实的基础平台。在市场竞争全球化和国内市场国际化的今天，谁能在技术创新上领先一步，谁就能在激烈的市场竞争中占据优势，谁就能够在今后的企业发展中获得先机。Angel CUE 系统的研发成功正是国内民航企业牢牢把握住“科学技术是第一生产力”的精神，将技术创新与企业发展紧密结合，在离港领域实现技术突破，从而创造出新的价值增长点，并为国内民航离港系统的进一步拓展奠定了坚实的基础，也同步提供了新的契机。

参考文献

- [1] IATA Recommended Practice 1797, Common Use Terminal Equipment
- [2] 中国民航信息网络股份有限公司 AngelCue 项目组. AngelCue 详细设计说明书 V1.3
- [3] UNISCOPE Protocol Reference Manual, Unisys Corporation, UP-10683 Rev.1
- [4] Ron Hitchens. Java NIO[M]. USA: O'Reilly Media. 2002. 120-123
- [5] RFC 2351, Mapping of Airline Reservation, Ticketing, and Messaging Traffic over IP
- [6] Herbert Schildt. Java: the complete reference[M]. USA: McGraw-Hill Osborne Media. 2005. 824-825

作者简介

王欣明（1975—），男，技术专家，中国民航信息网络股份有限公司研发中心产品开发部经理，主要从事民航信息系统的研究。

高新（1977—），女，工程师，中国民航信息网络股份有限公司研发中心产品开发部项目经理，主要从事民航信息系统的研究。

Research on Pork Safety Traceability System Based On RFID

TONG Xin-shun WU Yi

(Zhengzhou University of Light Industry, Economics and Management College,
Zhengzhou, Henan 450002)

Abstract: In the past 20 years, With the development of animal nutrition and feed industry, China's pork production has increased rapidly, But it also brought a lot of problems ,such as pork products supply system are complex, veterinary drugs and feed additives are abuse and so on,which affect the quality and safety of pork products.

The pork tracibility system can achieve the pork safety tracking and control throughout the supply chain by the of the information technology . Faced the Pig epidemic-prone, Pork security incidents continue and the requirements of the traceability for the international community of animal and animal product are urgent, to Improve the quality and safety of China's pork, improve the competitiveness of the pork market,make people can eat safe meat, it is necessary to establish a retrospective system for pork supply chain by the advanced technology.

Keywords: RFID; Pork supply chain; Tranceability

1 Introduction

Pig industry is an important industry in china , Pig product have a very important position in China's livestock industry. Pig industry is the main body of China's animal husbandry, and it is the main source of meat products. In the past 20 years, With the development of animal nutrition and feed industry, China's pork production has increased rapidly, But it also brought a lot of problems ,such as pork products supply system is complex, veterinary drugs and feed additives are abuse and so on.,which affect the quality and safety of pork products.

The pork tracibility system can achieve the pork tracking and control it throughout the supply chain by information technology . Faced with the pig epidemic-prone, Pork security incidents continue and the requirements of the traceability for the animal and animal product, to improve the competitiveness of the pork market,make the people can eat safe meat, establishing the tracebility system for the safety is necessary,which can improve the quality and safety of China's pork, However,the construction of pork trancibility system is a complex project, there are many links "from source to table", To achieve the control of tth quality of pork products, it is necessary to establish a retrospective system for pork supply chain by the advanced technology.

2 RadioFrequencyIdentification Technology (RFID)

The basic principle of Radio Frequency Identification (RFID) is electromagnetic theory. The advantage of RFID system is not limited to line of sight, the distande of Identification is farther than the optical system, RFID tags have a read-write capabilities, can carry large amounts of data, which are difficult to forge, and have a higher intelligence.

In general, radio frequency identification system contains radio frequency tags, readers and data management system^[1].

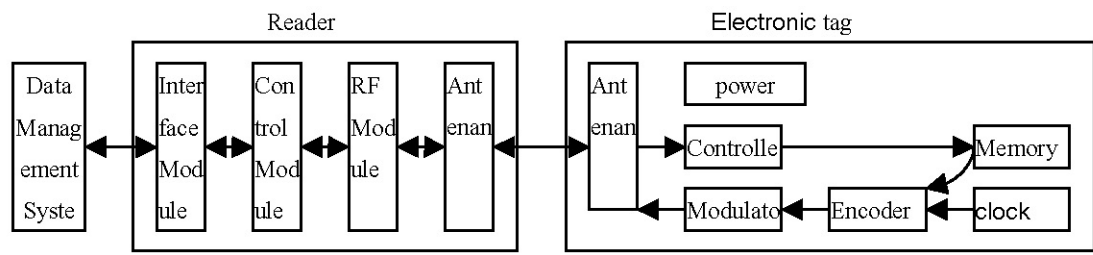


Figure 2-1 RFID identification technology diagram

Radio frequency identification technology has the advantages of its quick scan, small size, easy to package, anti-pollution ability and durability, reusable, penetrating and non-barrier reading, the data memory capacity, security, which are already used in the pork supply chain security identification extensively.

3 How the RFID technology are used in the pork safety trancibility system

3.1 Analysis of pork Business Process

There are a series of link in the pork business from the breeding, slaughtering and selling through a series of links. It is the flow of the process, including transport, storage and other links. Whose Circulation process contains the links of transport, storage, and so on. the extent of metamorphism is not only time-related, but also the environment of transport and storage sectors related, Such as temperature, humidity, illumination, ventilation conditions. It is necessary to trance these links in order to eliminate risks of pork^[2].

Through the analysis of the problems of existing systems, we set the methods for assessing security and trancing the pork by the technology of RFID to make sure the safety from the "farm to table". Its business processes as follows.

- (1) Pig farming is the source of pork, At this stage, we record the information of veterinary medicine, feed, breeding environment via by electronic tags.
- (2) There are a lot of integrated multi-sensor reader devices when pigs are transported from farm to the slaughterhouse arrangement which can record real-time information about this bichth pigs.

(3) the reader installed in the door will read the information of electronic tag on pork containers after from time to time when the pork transpted to the ogistics warehouse after slguchted.and sent to food safety management system Along with the sensor information. Besides the reader can read the time of storage, allocates inventory area by the system automatically. Warehouse is also furnished with integrated readers to read the tag a certain time interval.which have the information of environment.

(4) According to the records of environmental information, Logistics Warehouse evaluation system will play a role, assess the pork in Storage automaticly, to determine expired foods and the delivery order. Which will change the traditional assessing methods of "first-in-first-out", expired pork should be delieved firstly.

(5)safety pork will be transported to consumers after a rigorous flow of process. In this way, whether in or on the shelf beside the table, Consumers can not only understand the information of pork breeding ground, the consumption of feed, use of drugs, transport temperature, slaughterhouse name, selling land and do on, but also certificated the pork by pork safety evaluation system, enjoy "safety Meat".

3.2 Solutions of pork traceability system

The system contains electronic tags, antenna, electronic tag reader, the sensors to detect the parameters of external environment , pork trancebility database. Pig's ear tag is an electronic label, which can be read and written. Figure foollowed shows the architecture diagram of the pork safety trancibility system on RFID^[3].

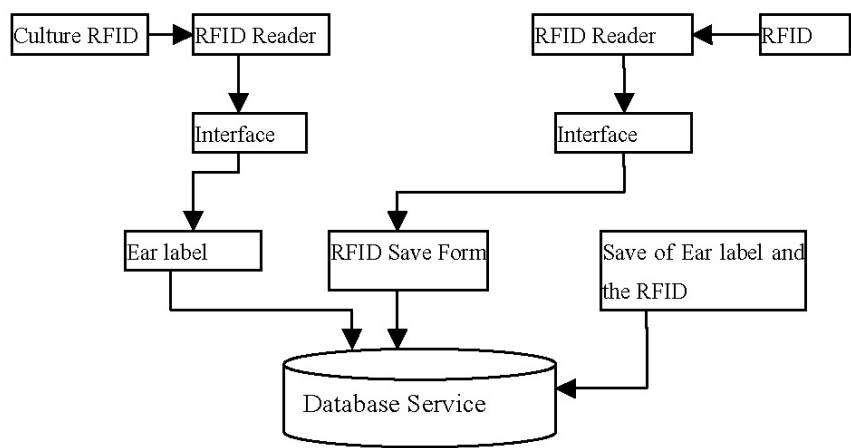


Figure 3-1 diagram of pork system architecture based on RFID

(1)In the stage of pig breeding , the reader operate the ear tag, Record the information of farms, feed, vaccines, veterinary drugs and so on.

(2) In the stage from breeding to slaughter, the Reader write the information of transport and environment.

(3) in the stage of slaughting, the reader read all of the information on RFID ear tags read

before the pig cutten the head

(4) after Ear tag information readed by the reader, the readed information are changed to the ear labels, which are sended to the computer by computer keyboard port

(5) the control program of the software system will delive the be pig ear label to the individual components in system software, which will manage the ear label in the database

(6) After the pig cutten, hanged a RFID on each piece of meat.

(7)Card information readed by the reader, the RFID information will be changed to the electronic identification number,which will be send to the computer by the computer keyboard port

(8) the program controled electronic identification number in Software system send the electronic identification number to the individual components, which will manage the ear label in the database

(9) When the quantity of pig completed slaughter, the procedures controled the electronic ID ear label and Electronic ID on Software systems will read the electronic ID ear label and Electronic ID respectively,and acheve the correspond of the electronic ID ear label and Electronic ID by the individual components

(10) When the meat refrigerated into cold storage , the reader re-read the information of each RF card, and changed to the information of pig's electronic identification data and send to the computer by computer keyboard port

(11)One-dimensional bar code printing control program passed electronic identification number to the individual components in system software, and the Individual component read pig ear label according to electronic identification number from the database.

(12) Converte the pig ear label to a one-dimensional bar code information, passed to the bar code printer, printing a one-dimensional bar code label, affixed to each piece of pork ,the control process of individual identity are completed.

4 The network structure of pork supply chain traceability system

In the traceability system, information systems have three distinct functions which are mainly acquisition, transmission and management . In the traceability system, the information of the attributes of the business product, and the participants is the basis ,the acquisition, transmission and management of information is the key to trancebility for food. we will study and explore the mothord of pork supply chain traceability system .

4.1 The developed network structure of traceability system of pork supply chain

Systems distributed in different regions, institutions and units, and they are usually connected by network. Usually the farms are away from urban, so they are lack of network technology. In general, farms are away from the city, and farming is the longest part in the supply chain, so the C/S structure is used. Farms themselves are responsible for the records of veterinary drugs, feed and the breeding environment, and the management of information data, including the pig birth, purchases

and sold record; The data is exchanged by desktop program and the pig archives are uploaded to the web server. The information of slaughterhouse and supermarket sales system is managed in local for a short time, while the application of network technology slaughterhouse and supermarket is relatively common, so the B/S (Borwes / sevice) structure is directly used, which meets the demand of customers for convenience and efficient queries. In the stage of slaughtering, the ear tag will be removed when the pig was cut, before which the data of breeding stage will be entered in the database. After the pigs are split in half, the tags are hung again, and the information of this stage will be written to RFID tags and entered in the database. This process involves the docking of the data in two stages, and then the tags are played. There is mainly query module in the sales subsystem, for the check of information in breeding, slaughtering and sales^[4].

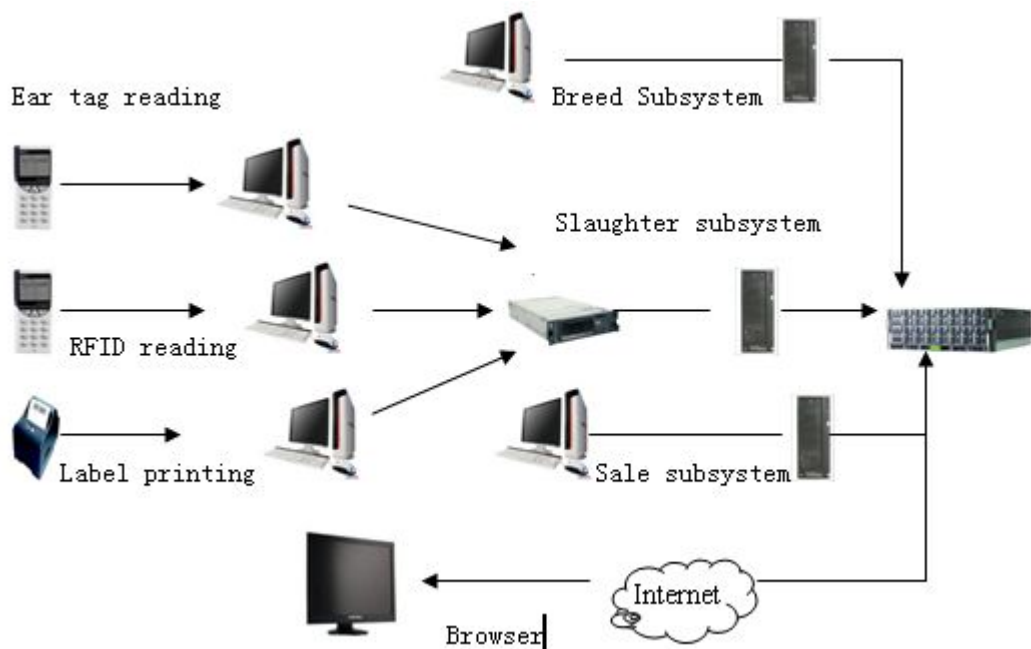


Figure 4-1 Pork supply chain traceability system network figure

4.2 the field analysis of System and the overall structure

Pork supply chain can be divided into three stages which are pig breeding, slaughtering and marketing. They are in different geographical realities, the phase varies greatly. This pork traceability system have three application modules.which are farm systems, slaughter systems and marketing systems^[5].

(1)The farm system

Pig breeding cycle is the longest link in the pork supply chain. We need four subsystems according to the analysis of Chapter 2,which are the standards and regulations subsystem, security and monitoring sub system, individual identity management subsystem, subsystem. he file management system. The Figure of pig farm system framework is as follows.

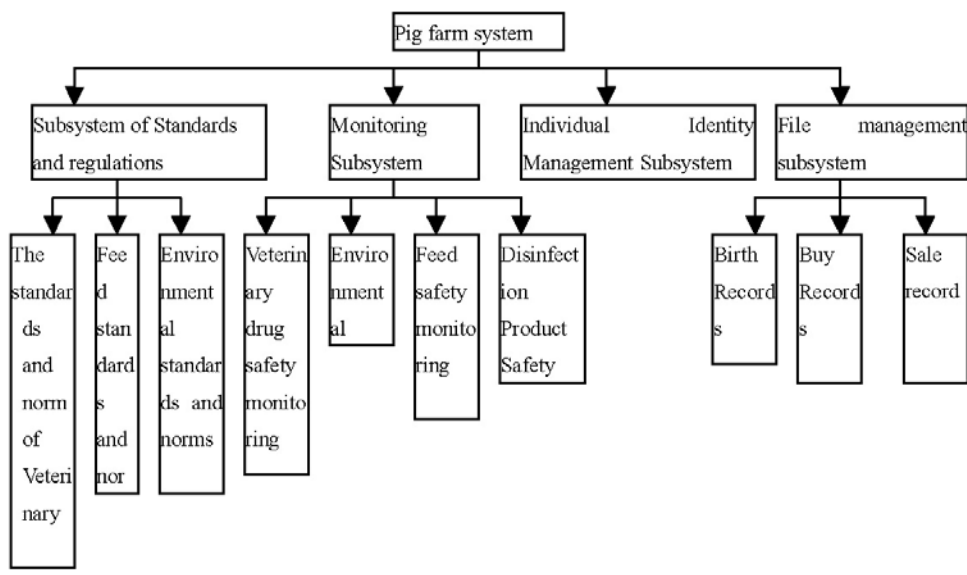


Figure 4-2 Farm system frame diagram

(2) Slaughterhouse system.

Slaughtering stage is the most complex aspect in the pork supply chain ,which is short and there are a lot of links. In Chapter 2, Slaughterhouse system includes four subsystems. File management subsystem includes records of pig transport, slaughter records, storage record of pork, pork transport records; safety monitoring system is mainly to monitor the quarantine before the pigs' slaughter, and record the information of the pork inspection, the slaughtering environmental monitoring, the transport disinfection, transport temperature and so on.The monitoring system provides the limits for the safety monitoring. Individual identity management subsystem mainly manage the electronic tags.

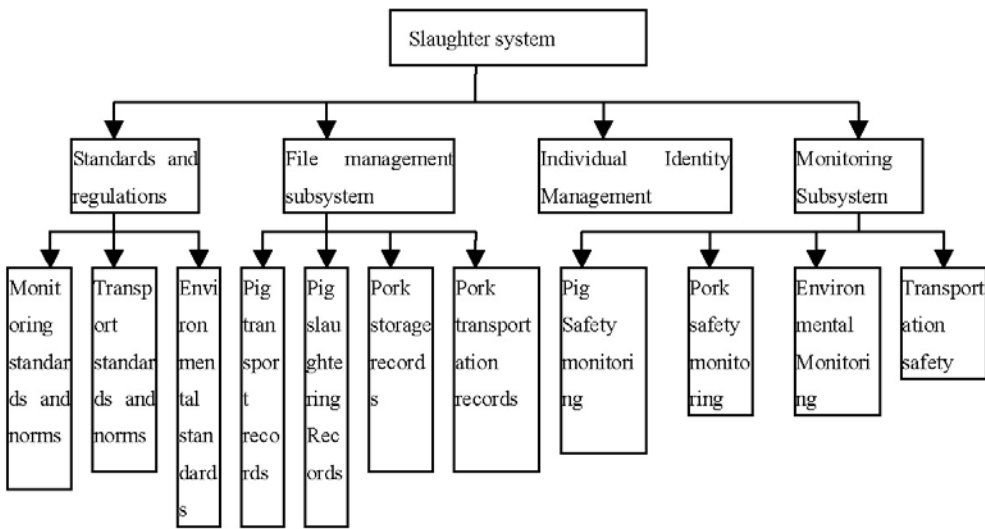


Figure 4-3 the Figure of slaughterhouse system

4.3 Marketing system

The trancibility of the Sales stage back mainly for the inquiry,the historical data can be recorded are not many. The marketing system can also be divided into four subsystems, Records management system is mainly related to pig and pork sales and storage, File query subsystem are mainly for pig breeding, slaughtering, sales recordss. Safety monitoring subsystem includes pork safety monitoring and environmental safety monitoring.

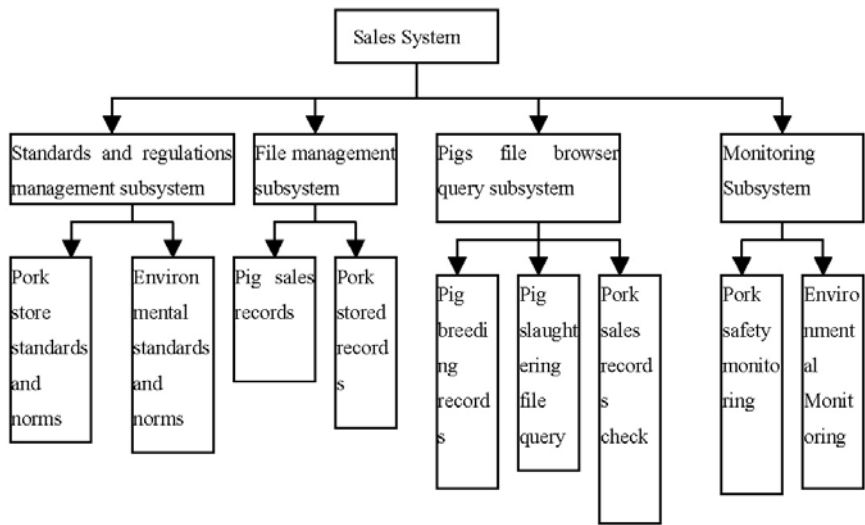


Figure 4-4 Sales System Chart

References

[1] Moe T.Perspeetives on traceability in food manufacture. Trends in Food science & Teehnology 9 (1998):211~214

[2] Jansen-VullersMH, CA van Dorp, A J M Beulens. Managing traceability in fomration In manufacture . International Journal of Information Management, 23(2003):395~413

[3] Blaneou J.A history of the traceability of animals and animal products.Revue Seientifique et Teehnique, 2001, 20(2):413~425

[4] Verbeke W. The emerging role of traceability and information in demand-oriented livestock produetion,Outlook on Agrieulture,2001,30(4):249-255

[5] Pascal Gerard,Mahe Sylvain. Identity, traceability and substantial equivalenceof food.Cellular & Moleeular Biology (Noisy-Le-rand),2001,47(8):1329~1342

Author

Tong-xinshun, Henan, Professor, direction: Logistics and Supply Chain Management

基于GTechnology的输电网WebGIS的设计与实现

徐雪荣 郭世界 王晓辉

(华北电力大学控制与计算机工程学院, 北京 102206)

摘 要: 根据电网实时监测管理的实际需求, 使用 GTechnology 作为地理信息系统 (GIS) 开发平台, 将先进的 AM/FM/GIS 技术、AJAX 技术、SVG 技术等应用到 WebGIS 开发中。采用面向对象技术建立设备模型, 构建了一个新型的集空间数据和属性数据与一体的输电网 GIS 数据库。通过与电力 SCADA 系统集成, 收集并分析设备实时数据。实现了矢量地图展示, 设备增删、设备属性查询和编辑, 图层显示控制, 监测设备实时数据展示等功能。

关键词: Gtechnology; 输电网; WebGIS; 设备模型; 实时

The Design and Realization of WebGIS in Transmission Network Based on GTechnology Platform

XU Xue-rong GUO Shi-jie WANG Xiao-hui

(School of Control and Computer Engineering , North China Electric Power University ,
Beijing 102206 , China)

Abstract: According to the actual needs of real-time power grid monitoring and management ,using GTechnology as the Geographic Information System (GIS) developing platform, the dvanced AM/FM/GIS 、AJAX and SVG technologies are adopted to develop WebGIS. Using the object-oriented technology,models of equipments are make up , a new-type GIS database of Transmission Network is builded by integrating spatial and attribute data. Accessing to Power SCADA System , the real-time data of equipments is collected and analyzed. Functions of Vector map showing , Equipment adding and deleting, Equipments' properties querying and editing, Layer display controling , Monitoring equipments' real-time data display are realized.

Keywords: GTechnology,Transmission Network,WebGIS,Equipment Model, real-time.

引言

近年来, 电力 GIS 系统开始成为电力企业的企业级空间图形资源平台, WebGIS 技术更是发展迅速。目前, 国内用于电力行业的 GIS 软件主要是 ESR 系列产品、Mapinfo 系列产品、

Smallworld、SICAD 以及国内的 Grow 等, 具有电力行业特色、功能强大的 GTechnology 软件在国内 GIS 市场中还没有得到很好的应用。

基于 GTechnology 平台的输电 WebGIS 是在上述背景下进行研究和开发的。系统的核心目标是建立企业级电网空间图形可视化管理平台, 分图层展示电网线路结构布局, 同时与电力 SCADA 系统集成, 分析、展示电网中的各个场站、设备、监测点的实时数据, 方便管理者以各种专题图的形式查看线路分布、为故障检修、线损分析、潮流计算、业务管理, 客户服务等业务提供依据。系统基于 B/S 模型, 通过在 GTechnology 和 GeoMedia WebMap 上做二次开发, 实现了输电网 WebGIS 系统。

1 开发平台和关键技术

1) GTechnology 平台概述

G/Technology 是美国 InterGraph(鹰图)公司的产品^[1]。G/Technology 作为在电力行业使用广泛的主要 GIS 平台之一, 与其他的平台相比, 主要有以下几个应用于输电网系统的优势:

- 面向电力行业;
- 开放性, 完全基于数据库的应用;
- 可扩展性, 面向设备对象, 遵循标准的体系结构;
- 高性能, DDC (动态显示缓存)+DELTA (数据记录集)。

2) GeoMedia WebMap 平台概述

GeoMedia WebMap 是一个完整的 Web GIS 开发环境, 采用 asp 实现, 提供了一整套用于 Web 开发的高级空间和网络分析的对象和服务供程序员使用, 不依赖于地图定义文件, 直接用代码操作这些对象生成网上地图^[4]。通过它用户可以创建动态的、自定义的、可对空间数据源修改的、在 Web 上用于对地理数据进行浏览和分析的空间应用系统。

3) 动态网页技术

由于 webGIS 是基于 GeoMedia WebMap 提供的服务开发的, 所以本系统选择 ASP 作为动态网页开发语言。

在对 GIS 进行 Web 发布时, 应用 AJAX 技术完成客户端脚本与服务器之间的数据交互过程, 实现网页动态的局部刷新。输电网 WebGIS 需要实时刷新电气监测设备状态数据, 而输电网地理信息不需要改变, AJAX 的异步对象调用技术正好满足了这一要求。AJAX 技术的优点在于, 它向开发者提供了一种从 Web 服务器检索数据而不必把用户当前正在观察的页面回馈给服务器。与现代浏览器的通过存取浏览器 DOM 结构的编程代码(Javascript)动态地改变被显示内容的支持相配合, AJAX 让开发者在浏览器端更新被显示的 HTML 内容而不必刷新页面。

4) SVG 技术

SVG (Scalable Vector Graphics) 是一种全新的矢量图形规范。由于 SVG 支持脚本语言 (script), 可以通过 Script 编程, 访问 SVG DOM 的元素和属性, 即可响应特定的事件, 从而提高了 SVG 的动态和交互性能。SVG 实现了图形、图像和文字的有机统一。这就为实现 WebGIS 提供了可能。

SVG 提供了丰富的图形对象, 包括直线、路径、圆、图标、文字、图像等, 满足了 GIS

系统的需要。GIS 系统最基本的功能是地图控制，SVG Viewer 本身提供图形的缩放功能。SVG 采用基于 XML 的 DOM 文档管理结构，很方便实现层次管理。SVG 图形数据本身只包含用来实现矢量图形显示的信息，如坐标点、变换矩阵、显示样式等信息，不能满足 GIS 系统的需要。但由于 SVG 是基于 XML 格式的，除了 使用其内置的属性外，可以对其属性进行任意扩充，以实现自定义的功能。

2 系统分析与设计

1) 系统体系结构

基于 GTechnology 平台的输电 WebGIS 采用 B/S 体系结构作为系统的总体解决方案，将 COM 组件技术和 ActiveX 技术分别应用在服务器端和客户端上。系统以 G/Technology 作为 GIS 开发平台，用设备模型来描述输电网络。以 GeoMedia WebMap 作为 WebGIS 开发平台，asp 作为程序开发语言，直接读取 DDC 空间数据文件在 IIS 中发布，实现在浏览器中生成 SVG 地图，展示实时数据，提供相应的操作。数据库采用 oracle9i 及以上版本。

基于 GTechnology 的输电网 WebGIS 主要由 Web 服务、GIS 服务和数据存储三部分的关键技术构成，采用基于 B/S 方式的三层体系结构，如图 1 所示。

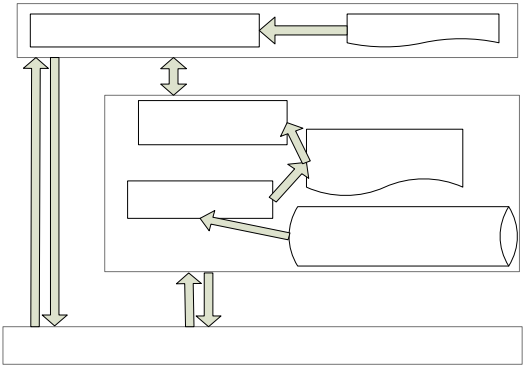


图 1 系统三层体系结构图

2) GIS 数据组织策略

空间数据库是 GIS 系统的核心。以往的 GIS 系统都是采用数据库加文件的存储策略，地理设施的属性数据存在数据库中，空间数据则以地图文件的形式存储。基于 GTechnology 的输电网 WebGIS 集空间数据与属性数据于一体，完全基于关系型数据库存储，GIS 数据完全开放，方便用户查看。

采用面向对象的技术，按照不同的线路电压等级和线路、设备类型建立设备对象模型。WebGIS 的每个图层对应一种设备，每种设备由属性组件，符号组件和标注组件组成。其中符号和标注是图形化的组件，属性是非图形化组件，每个组件对应数据库中的一张表，这样就可集中管理设备的空间数据与属性数据。为了能清晰的区分出电气设备的正常和故障情况，系统中为每个设备建立相应的报警规则，同时设定满足不同规则时的地图显示样式，比如改变设备符号颜色，高亮显示等。这些定义存储在每个设备对应的样式定义表或样式规则表中。

同时接入电网 SCADA 系统数据库，收集电网中各监测设备的实时数据，并在地图上对

Web 服务器

应设备位置处展示。

3 系统功能实现

1) 基于 GTechnology 的 GIS 开发

系统以 GTechnology 软件作为基础的 GIS 开发平台，主要包括如下几个步骤，在 GTechnology 中都有友好的引导界面辅助用户操作。

① 建立设备模型：抽象每种设备的属性信息和图形信息，建立设备的各组件表和显示样式表，结合 GIS 符号系统设定设备的显示样式和规则。

① 导入 shape 格式的矢量地图数据：采用山川、河流、居民地等自然地理矢量图作为输电网的背景图，为输电设备提供位置参照信息。

② 描绘地理接线图：导入 GPS 采集来的杆塔、变电站等设备数据后，确定各杆塔之间的电气连接关系，设置各线路的属性信息，完成数据库的更新。

2) GIS 数据的 Web 发布

① 输电网矢量图形展示

实现基于浏览器的 GIS 系统，需要将 SVG 图形对象嵌入到网页中，使用 HTML 代码来实现。浏览器回调用 SVG Viewer 在指定区域显示图形。实现地理接线图展示，图形放大、缩小、漫游、查看信息等，并显示当时比例尺的大小。编程开发的流程图如图 2 所示。

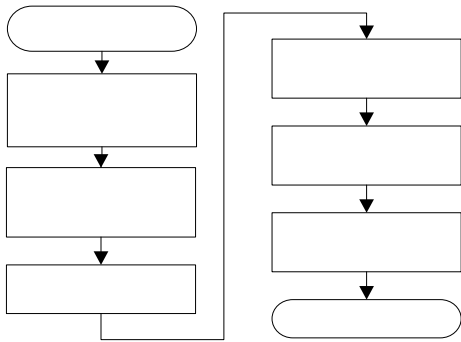


图 2 矢量图形展示开发流程图

② 设备增删和修改、设备属性查询和编辑

在 SVG 图形中，用编号 id 来唯一标识一个设备对象，即可通过 SVG 文档对象的 getElementById()函数来获取指定的设备对象，进行指定操作。

③ 图层显示控制功能

每个图层存储一种设备类型，通过采用 SVG 的组对象来实现图层显示控制，不同图层的对象

包含在不同的组中。比如<g id="杆塔">< style = " visibility : visible ;color=blue;" /> </g>，id 表示定义的是杆塔设备，属性 visibility 设置可见性，color 设置颜色，还可增加属性设置选中、删除所有对象等操作。

④ 监测设备实时信息显示

实时数据是由电力 SCADA 系统接入，程序中采用了 AJAX 的 XMLHttpRequest 对象来实现动态的刷新实时数据。一个经由 XMLHttpRequest 对象发送的 HTTP 请求并不要求页面中拥有或回寄一个<form>元素。XMLHttpRequest 对象的 send()方法可以立即返回，从而让 Web 页面上的其他 HTML/JavaScript 继续其浏览器端处理而由服务器处理 HTTP 请求并发送响应。

4 应用实例与分析

上述设计方案已被应用到实际开发中，下面以山西省晋中市的电网实时监测管理 WebGIS 为例，展示采用上述方案开发的系统效果，如图 3 所示，是通过图层控制显示的污区图，图 4 所示是线路、杆塔分布及监测点实时信息。

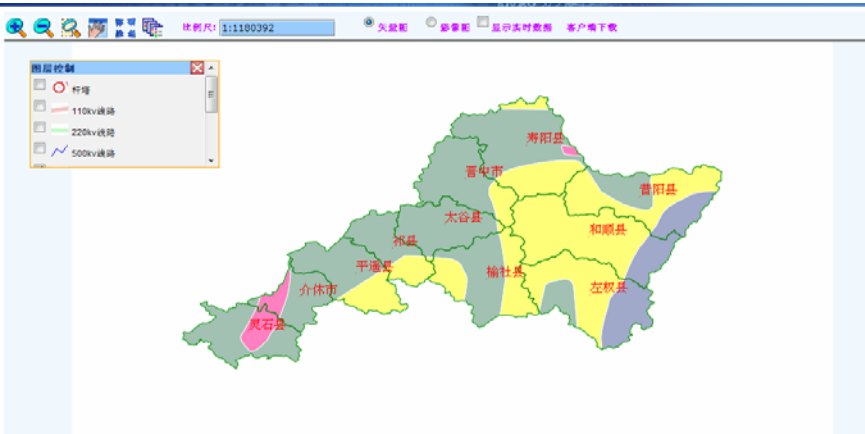


图 3 污区图

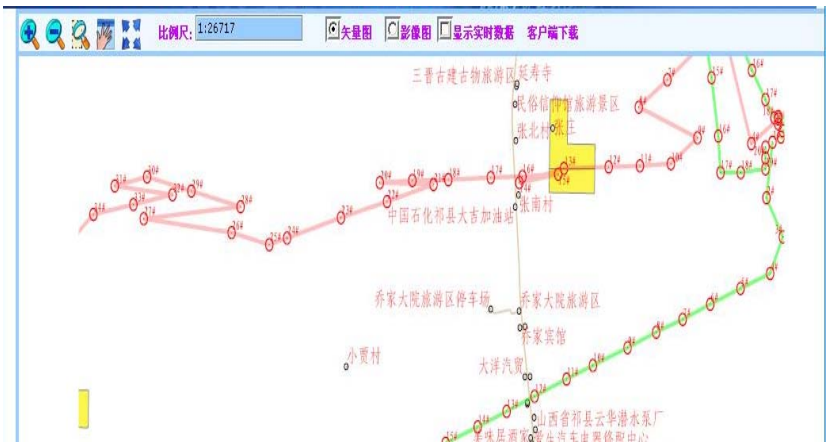


图 4 信息展示

实际应用表明，采用 GTechnology 作为开发平台，开发电力 GIS 流程简单，扩展和维护方便，系统具有电力行业特色；将 Ajax、SVG 等技术应用到 WebGIS 中，很好地提高了系统性能，减少不必要的延迟，带来了很好的视觉效果，满足监测、管理电网的要求。

5 结束语

通过 WebGIS 技术在电网管理中的应用可以方便、直观地监控全局运行情况，并为输电线路实时监管系统、潮流计算系统、线路故障检修优化系统等提供可视化展示平台。实践证明，基于 GTechnology 平台的输电网 WebGIS 符合电力 GIS 要求，已被应用到山西省电力系统的实际项目开发中，功能完善，性能良好。

参考文献

- [1] 高铁军. ArcGIS/GTechnology 平台在城市管网地理信息系统中的选择[J].测绘科学, 2008, 33:173-174.
- [2] 李景文, 周文婷, 刘军锋. 基于地理实体的面向对象矢量模型设计[J].地理与地理信息科学,2008, 4: 29-31.
- [3] 马甲军, 杜玉蕾. 电力 GIS 中数据模型的选择[J].电测与仪表,2009, 6: 56-58.
- [4] 廖兴, 朱砚. 贵州电网输电 GIS 系统的构建与实现[J].广东输电与变电技术,2009, 11(2): 19-22.
- [5] 任怀兵. 青藏铁路输变电 GIS 设计与实现[D].西南交通大学研究生学位论文: 西南交通大学, 2003.

作者简介

徐雪荣（1987—），女，硕士，河南人，华北电力大学控制与计算机工程学院硕士，研究方向为电力智能软件技术。

郭世界（1971—），男，硕士，山东人，研究方向为电力信息技术。

王晓辉（1984—），男，博士，山东人，华北电力大学控制与计算机工程学院博士，研究方向为电力智能软件技术。

虚拟现实中的力/触觉建模技术

徐玉彬 刘玉庆 朱秀庆

(中国航天员科研训练中心, 北京 100094)

摘 要: 本文着眼于我国航天事业的发展, 总结了现有的力触/觉建模技术, 介绍了具有力/触觉建模的虚拟现实技术在实际中的应用, 并提出力/触觉建模技术在我国航天员虚拟训练中的应用前景。

关键词: 虚拟现实; 力/触觉建模; 航天员训练

The Technology of Haptic Modeling in Vitual Reality

XU Yu-bin LIU Yu-qing ZHU Xiu-qing

(China Astronaut Research and Training Center, 100094 BeiJing)

Abstract: This paper is based on the development of Chinese manned spaceflight Cause. The overview of the haptic modeling technology and its application are provided. The application of VR with haptic feedback for China astronaut virtual training is discussed, and the research direction of haptic modeling for astronaut training is given by the end of this paper.

Keywords: virtual reality; haptic modeling; astronaut training.

1 引言

随着我国载人航天事业的发展, 各种空间试验任务逐渐增多, 航天员也将面临日益增多的出舱活动任务。出舱活动是风险最高的航天任务之一, 着舱外航天服的航天员活动受限, 操作任务复杂, 对航天员在太空失重环境下个人操作技能要求越来越高, 对地面训练仿真技术手段也提出了更高要求, 现有的训练设备在很多方面具有较大的局限性。从国外航天员训练相关科研领域的发展来看, 虚拟现实技术作为一种高效、可行的技术手段, 已广泛应用于航天员技能训练。

在虚拟手交互过程中, 要让用户获得“身临其境”的逼真感觉, 仅有手的视觉信息, 而没有手的力觉感知是远远不够的。例如, 在运用虚拟交互技术进行拆卸和传递负荷训练时, 如果航天员只在视觉上看到载荷, 却感受不到拆卸它和传递它时手上受的力。那么航天员就难以对出舱过程有一个完整认识, 也就达不到模拟太空环境进行仿真训练的要求了。为了增强航天员出舱仿真训练的沉浸感和真实性, 我们渴望航天员在看见虚拟手抓取物体的同时, 能够感受到真实的作用力, 即能感受到与现实世界一样的力。若航天员感受到的力不真实、

不准确,那么他所做出的决策将不可避免地会发生错误,从而难以实现对虚拟物体的正确操作,同样也使得虚拟手交互的真实性大打折扣。

2 国内外的研究现状

力/触觉建模研究最早可追溯到 20 世纪 50 年代,在远程控制的机器人系统中,德国科学家开展了力/触觉生成研究。20 世纪 90 年代初以来,力/触觉的生成和反馈逐渐成为人机交互领域研究的热点,国内外许多科研院所开展了大量研究^{[1][2]}。

随后,国内外许多专家和学者开展了多个方向、不同角度的研究,改进或提出多种“haptic”生成算法。

目前,最常见的触觉和力觉生成算法是利用 PHANTOM 设备,根据虚拟物体的刺穿深度,运用虎克定律计算接触力。当用户手指操作 PHANTOM 设备与虚拟环境交互,作为手指的化身“avatar”接触到虚拟物体时,运用虎克定律计算作用力,并且反馈到用户的真实手指上。

国内对虚拟力建模也开展了研究。研究单位包括:浙江大学 CAD&CG 国家重点实验室、东南大学、北京航空航天大学、中国科学院软件研究所、华中科技大学等。

浙江大学杨文珍博士提出一种基于物理的虚拟手交互真实力觉生成方案^[1]。在其论文中根据虚拟手的位置信息的改变来计算出虚拟手运动的速度和加速度,然后利用牛顿第二定律和物体本身的质量参数计算出虚拟力的大小。该方法能够进行全面的受力分析,真实的计算出力的大小,反应速度也能满足训练的要求,值得借鉴。

东南大学吴涓和宋爱国学者提出了一种基于物理意义的快速力反馈形变模型及实时力觉响应算法^{[3][4]}。该方法运用基于物理意义的形变对柔性体进行建模,不仅计算速度快,满足力反馈的实时性要求,而且能够同时保证接触力和形变的计算具有较高的精度,满足精细作业对虚拟现实系统的要求,并且可以处理各向异性的情况。该方法为虚拟手操作柔性物体的力觉生成问题提供了良好的解决途径,值得借鉴。

浙江大学朱振华根据构造出的虚拟手抓取轨迹线进行碰撞检测,然后利用基于虚拟物体刺穿深度的算法计算虚拟力的大小^[5]。该方法的优点是能够快速计算出力的大满足实时性的要求,易于算法的实现,不足之处在于力的真实性有待提高。

3 力/触觉建模技术

触觉与力觉是有区别的。一是力觉反馈,用来感受虚拟物体的重量、惯性、硬度、粘性、运动约束等;二是触感反馈,用来感受虚拟物体表面粗糙度、几何形状、温度、材质、湿度等。我们在这里讲的力触觉建模技术其实主要是指力觉的建模技术。

力/触觉建模技术是指利用基于物理的或非物理的分析方法,建立一种能够计算出虚拟环境中虚拟手与虚拟物体之间进行交互产生的相互作用力大小的模型。力/触觉建模技术还可以分为对与刚性体的和对柔性体的建模技术。

3.1 虚拟力分类

虚拟力可以分为接触力和碰撞力。接触力是虚拟手与虚拟物体表面接触时手指感受得到作用力,例如手抓着物体运动或手握杆攀爬,都是属于接触力得范畴;碰撞力是指两个或两个以上相对运动的物体接触并伴有速度突然变化的现象的过程中所产生的力。由于碰撞的时间间隔极其短促,碰撞后物体速度发生有限的突然变化,所以物体间作用的碰撞力很大,使物体发生变形,且通常伴随着发声、发热、甚至发光等物理现象。碰撞的实质在于相碰物体的弹性及塑性变形作用,可将物体碰撞过程可分为压缩变形阶段和弹性恢复阶段。例如日常生活中的钉钉子、乒乓球拍击球过程中的作用力都属于碰撞力。

3.2 基于物理的力觉/触觉建模技术

基于物理的力觉/触觉建模技术^[6]指的是将更多的物理属性赋予虚拟物体,如质量、惯性、摩擦、阻尼等,引入到虚拟力的计算模型中,生成力觉。根据虚拟手与虚拟物体的多个接触点处的受力进行分析,遵从力是矢量、力平衡、力矩平衡等基本力学规律进行求解,计算出虚拟物体在多个接触点处的接触力、阻尼力与摩擦力并作为最终的作用力反馈给用户。这种方法从物理属性进行分析,遵循力学客观规律,生成的虚拟力具有较强的真实感与可靠性。

3.3 基于传感器的力觉/触觉建模技术

基于传感器的力觉触觉建模技术指的是把传感器测得的力的数据和通过其他方法同时测得的人体的其他参数,例如接触面积、运动加速度、关节角度等进行拟合得到一个经验公式,或者是把它们建立成一一对应的表格。在力觉计算的时候我们就可以把接触面积、加速度、关节角度等作为输入来进行计算或者查表得到力的大小。

杨文珍在他的博士论文里提出了一种利用压力传感器测量手指接触力的实验方法。他根据手指接触变形和相应接触力大小之间的关系,提炼出手指接触面积和接触力之间的力学模型,用于手指接触力的计算^[1]。

3.4 基于肌电信号的力觉测量技术

现在在国际上利用肌电信号测力是比较前沿的研究。它是指利用采集到的人体手臂的肌电信号通过时域、频域和小波分析的方法来来对手臂或手指用力的大小进行分析。重庆大学的侯文生教授利用肌电信号分析^[7],得到了桡侧腕长伸肌的积分肌电图与人手的握力大小成正相关性的结论。从理论上来讲利用肌电信号测力可用行得通,但目前还停留在实验室阶段,若想用于实际测量应用还有很长一段路要走。目前肌电信号主要应用于机械手的控制、假肢控制、疲劳度分析、疾病康复训练等方面。

4 力/触觉人机交互接口装置

力/触觉反馈设备是实现虚拟现实系统中力/触觉人机交互的必备装置。计算机触觉领域的

研究始于触\觉反馈的硬件研发，并随着硬件设备性能的不断完善和新设备的涌现，得以快速的发展。

力/触觉反馈设备种类繁多，功能各异，大小和用途也都不尽相同。虽然可以根据机构设计、动力性能、人机工程、用户交互部位和研发单位等对其进行分类，但想详细的分类也并非易事。下面是几种比较典型的力反馈装置^{[8][9]}。

图 4.1 所示的 CyberGraspTM 力反馈外骨架是美国 Immersion 公司研发的背置式多指力反馈设备，应用较广。它是由电机牵引丝线来带动手指，给手指施加一定的力反馈，力的范围是 0~12N，这是一种主动式的力反馈设备。

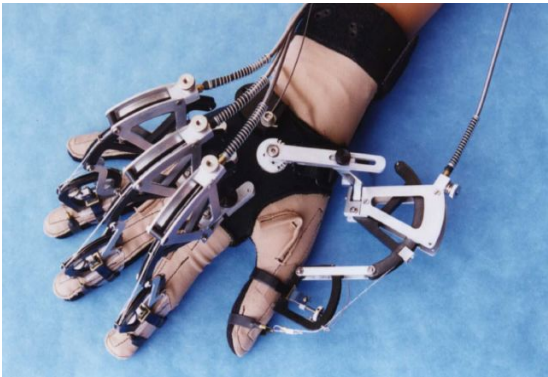


图 4.1 CyberGraspTM 力反馈外骨架

图 4.2 所示 Rutger Master II-ND 是美国 Rutger 大学研发的一种内置式多指力反馈设备，它是利用气体泵来产生力觉反馈，是属于被动式的力反馈装置，通过改变气体泵的压力大小来调节对于人手指力反馈的大小。



图 4.2 Rutger Master II-ND

图 4.3 为某种型号的 Phantom 设备。由美国 SensAble Technology 公司生产的“Phantom” Personal Haptic interface mechanism 力反馈设备是目前应用最广泛的 haptic 设备。当用户的手指套在 Phantom 的指环上与虚拟物体接触时，三个直流电机牵拉引线，在指环上作用 X，Y，Z 坐标上的三个方向的力。

图 4.4 所示是由东南大学研制的磁流变液力反馈装置。它根据磁流变液可以随着磁场变化

而改变状态的特性来改变对手指的阻力，这也是一种被动式的力反馈装置。



图 4.3 Phantom 设备



图 4.4 磁流变液力反馈数据手套

图 4.5 所示为微软公司的力反馈方向盘，它被用于各种赛车游戏，玩家可以通过这种力反馈装置可以体验到真实的驾驶体验，获得更多的快感。



图 4.5 微软力反馈方向盘

图 4.6 所示为微软公司开发的一款具有力反馈的摇杆装置，它也被广泛的应用与各种游戏，它会随着游戏的情节来调节内部的电机来给予玩家不同的刺激。



图 4.6 微软力反馈摇杆

力反馈设备经过这么多年的发展已经有了很多种的形式，也涉及各个方面，但是他们也都存在一些共同的缺陷。第一，现有的力反馈设备的自由度比较少，只能对与某些固定的动作进行反馈，对于复杂动作就无能为力。第二，现有的力反馈设备只能单纯的反馈力的感觉，并不能反映出物体的材质或者柔软度等触觉信息，有待进一步的发展改进。相信在不久的将来会有更多的新型的力反馈装置涌现出来以满足人们日益增长的需求。

5 力/触觉技术在航天员训练中的应用

5.1 航天员手部操作动作的分类

神舟七号的发射成功和中国首次太空行走的完美实现标志着我国航天事业达到了一个全新的高度。随着我国航天事业的发展和今后空间站的建立，我国航天员将会面临更多的空间操作任务，为保证航天任务的顺利完成，航天员地面训练是载人航天不可缺少的重要环节。将力/触觉技术应用到航天员空间操作地面训练中是力/触觉技术应用的一个重要方向。

经过分析可以把航天员在神七任务中的手部操作动作分为以下两类：跨握、圆柱抓握^[10]。

跨握：主要是指航天员在太空中抓持载荷运动，其基本手型如图 5.1 所示。拇指和其他手指分开，分别位于被抓持物体的两侧。由于处于失重状态，所以航天员手部对载荷所施加的力全部用来改变载荷的姿态或运动状态。

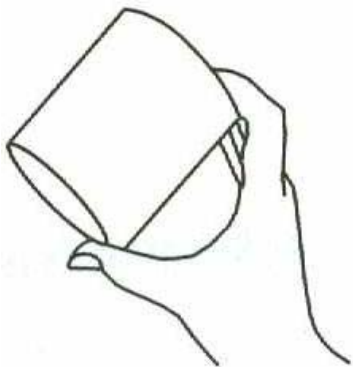


图 5.1 跨握示意图

圆柱抓握：主要是指航天员舱外行走手抓扶手的动作，手型如图 5.2 所示。

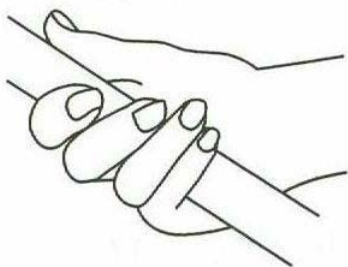


图 5.2 圆柱抓握示意图

在出舱活动中，航天员凭借圆柱抓取动作实现在舱外沿扶栏的太空行走。事实上，航天员在太空行走过程中，其身体受力情况是非常复杂的，而这所有复杂的力都来自航天员双手作用在扶栏上以后，所产生的反作用力。另一方面，圆柱抓取本身也是一个非常复杂的手部动作，其复杂之处在于，手（包括手指和手掌）与所抓取的圆柱体之间形成一个紧密的圆弧型接触面，对这个接触面，既不能按着点接触假设下的各种物理模型进行分解，也无法使用二指抓持假设实现力的合成。

5.2 手部操作虚拟力计算

失重环境下跨握力的计算模型的建立是虚拟力建模一个重点也是难点，为了达到仿真训练的沉浸感与真实性，必须要快速、准确地计算出航天员操作时各手指所受力的的大小。综合分析了各种前人的研究方法，决定采用基于物理的受力分析方法来实现这一目标。

航天员进行舱外行走时手型为圆柱抓握，手紧抓扶手，通过和扶手之间的作用力与反作用力来推动自己运动。由于手和扶手之间的接触方式比较特殊，无法利用点接触的方式来描述，因此无法利用力螺旋平衡的方法来求解航天员手部的受力大小。目前拟采用模拟失重实验的方式得到航天员在操作过程中的手部受力情况和航天员运动中的手指关节角度信息以及身体质心加速度信息，以此来找出它们之间的联系，对航天员圆柱抓握时手部受力进行建模。

5.3 力觉交互回路的实现^{[11][12]}

力觉交互回路是指用户通过触觉和力觉反馈硬件设备感知虚拟环境，通过力/触觉反馈设备来和虚拟环境进行具有真实力感的交互操作。只有在虚拟现实系统中引入触觉与力反馈，才能真正建立一个“看得见摸得着”的虚拟环境。

力觉交互回路主要由用户、触/力觉反馈硬件设备、触/力觉处理器和主机组成。触/力觉反馈硬件设备的主要功能是利用传感器测量用户的运动和位置，且能将虚拟环境中生成的力感或触感反馈给用户。触/力觉处理器用来处理位置、运动、触/力觉等数据，并且与主机进行信息传递。

回路输入：数据手套输入航天员手指的关节角度。位置跟踪仪反馈人手抓持载荷运动的位置变化信息。

回路输出：各个手指力的大小，通过力反馈设备把力觉反馈到航天员的各个手指上。为

虚拟操作训练提供力觉反馈。

6 结束语

本文介绍了虚拟现实中的力触建模技术，对目前存在的几种力触建模方法进行了综述，并给出将力触建模技术及力反馈技术引入到我国航天员虚拟训练中的基本思路，对我国航天员虚拟训练具有十分重要的应用价值。

参考文献

- [1] 杨文珍, (2006). “虚拟环境中具有力反馈的灵巧手抓取方法研究及应用”, 浙江大学博士学位论文
- [2] 汪成为 (1996). 灵境(虚拟现实)技术的理论、实现及应用. 北京, 清华大学出版社.
- [3] 吴涓, 宋爱国, 李建清 (2005). "一种基于物理意义的快速力反馈形变模型及实时力觉响应算法." 传感技术学报 18(1): 90-94.
- [4] 陈旭, 宋爱国, 李建清 (2006). "一种新的虚拟纹理触觉再现方法及其装置实现." 测控技术 25(8): 72-75.
- [5] 朱振华, (2006) “虚拟环境中具有力反馈的灵巧手抓取方法研究及应用” 浙江大学硕士学位论文
- [6] 杨文珍, 高曙明, 万华根, 朱振华, 骆阳 (2005). "基于物理的虚拟手抓持力觉生成和反馈." 计算机学报 28(6): 959-964.
- [7] 吴小鹰, 侯文生等 (2008) “腕长伸肌表面肌电与握力大小的相关性研究” 仪器仪表学报
- [8] 童明, (2001). "面向虚拟制造和装配的触感接口." 机械设计与制造工程 30(6): 24-26.
- [9] 董士海(2004). "人机交互的进展及面临的挑战." 计算机辅助设计与图形学学报 16(1): 1-13.
- [10] 张玉茹 等, (2007) “机器人灵巧手——建模、规划与仿真” [M]. 北京: 机械工业出版社, p.40
- [11] 刘杰, (2004) “虚拟环境中灵巧手主从抓持的实现.” 机器人 26(2): 107-110.
- [12] 陈卫东, 席, 蔡鹤皋 (1999) “虚拟现实系统中的手部跟踪和力觉再现技术研究.” 测控技术 18(6): 18-21.

作者简介

徐玉彬 (1985—), 男 (汉族), 河南省开封市人, 硕士研究生, 主要研究领域为人机环境工程学。

刘玉庆 (1962—), 女 (汉族), 天津市人, 研究员, 硕士研究生导师, 主要研究领域为航天飞行训练仿真技术。

朱秀庆 (1964—), 男 (汉族), 北京市人, 副研究员, 硕士研究生导师, 主要研究领域为航天飞行训练仿真技术。

古建动画自动生成系统中构件位置计算

尹梅芳 朱宫瑾

(北京工业大学计算机学院, 北京市多媒体与智能软件重点实验室, 北京 100124
中国科学院数学与系统科学研究院, MADIS, 北京 100190)

摘 要: 古建动画自动生成系统根据用户对建筑结构的描述、自动生成三维动画来表现古建的搭建过程, 大木构架核算模块是其中的一个重要模块, 该模块以用受限自然语言描写的古建筑制式为分析对象, 通过一系列的分析、推理、计算, 得到构成该古建的所有构件的信息, 本文所述的构件位置计算是大木构件核算的中间环节, 本文在学习总结古建筑建造规则的基础上, 提出了计算机自动计算古建筑各构件在三维场景中位置的方法。

关键字: 动画自动生成; 人工智能; 中国古典建筑

1 引言

90 年代, 中科院陆汝钤院士提出了全过程计算机辅助动画自动生成技术^[1, 2], 将人工智能技术和基于知识的方法引入动画生成的全过程。该技术将故事以受限自然语言的形式输入计算机, 直到最终生成动画, 每一步都是在计算机的辅助下完成的。

中国古典建筑是中华文化的瑰宝, 现存的古典建筑所体现出来文化内涵是五千年历史的积淀。随着计算机产业的发展, 计算机技术已经应用到我们生活中的方方面面, 但中国古典建筑信息化方面所做的研究却少之又少。本项目组在计算机辅助动画自动生成技术研究的基础上, 将这一技术应用到古典建筑大木作结构的研究中来, 做出了计算机技术应用到中国古典建筑领域的初步探索, 对于我国物质文化遗产的保护具有重要意义。

古建动画自动系统是基于这样的设计理念实现的: 该系统将一个合适的中国古典建筑描述以受限自然语言的形式输入到计算机, 经过大木构架核算、构件及人物动作规划、建筑场景规划等步骤, 最终生成动画片, 这些过程一部分是计算机独立完成, 一部分是在计算机辅助下完成的。

本文所述的构件位置计算是大木构架核算一个中间环节。大木构件核算模块是计算机独立完成的模块。大木构件核算模块以用受限自然语言描写的古建筑制式为分析对象, 通过一系列的分析、推理、计算, 得到构成该古建的所有构件的信息, 将这些信息用专业的 3D 软件读取, 可以看到三维场景中一个完整的古建, 实现了从平面语言到立体场景的实时转换。

本文第二部分介绍了构件位置计算所属模块—大木构架核算模块, 第三部分介绍了构件位置计算的设计思想及实现方法, 第四部分总结了古建动画自动生成系统的对于信息化古建筑研究和保护工作的重要意义。

2 古建木构架核算模块

1) 整体描述

大木构架核算模块是古建动画自动生成系统的第一个环节，大木构架模块的输入是用受限自然语言描述的古建筑的制式，如 N 间 M 檩，庑殿建筑。输出是包含所有构件的位置、旋转角度、尺寸、建造顺序的文件。从输入到输出要经过一系列的分析、推理、计算，首先根据输入信息分析出所有构件的名称及每个构件的建造顺序；然后，构件的名称又作为下层的输入，通过尺寸计算模块计算出所有构件尺寸；最后位置计算以上层计算和推理的结果为输入，构件名称和构件尺寸又作为位置计算的输入，计算出所有构件的位置、旋转信息。这三部分的推理分析都是建立在知识库的基础上进行的。

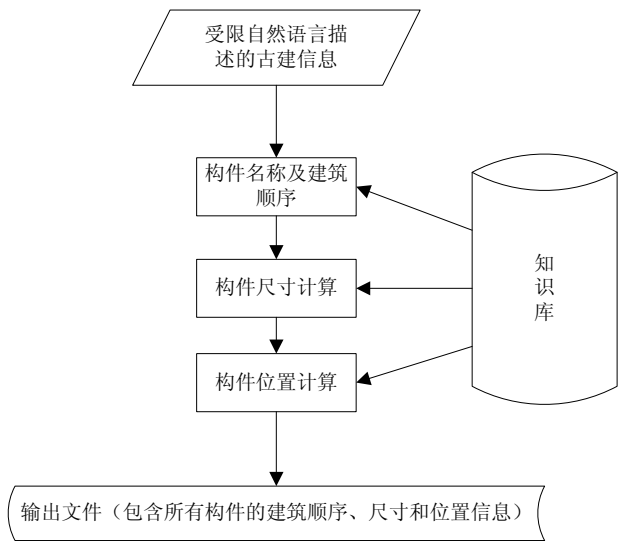


图 1 木构架核算模块

2) 研究范围

中国古典建筑经历了几千年的演变，涵盖了多种不同的文化元素，其建造规则也多种多样。明清建筑离我们比较近，遗存下来的实物较多，古建筑域这方面的研究成果也较丰富，古建领域的规则较统一。所以本系统以明清建筑为主要研究对象，知识库的构建及计算程序的设计主要以明清建筑的构造法则为基础。

清代大式建筑分为四大类：硬山建筑、歇山建筑、悬山建筑和庑殿建筑^[3]，其中硬山建筑是最简单的清代建筑，庑殿是最复杂的、代表皇权的最高制式的建筑。因悬山建筑与硬山建筑在大木结构上基本相同，本系统主要以硬山建筑、歇山建筑和庑殿建筑的构造规则为基础，实现这三类建筑多种制式的自动生成。

3 构件位置计算

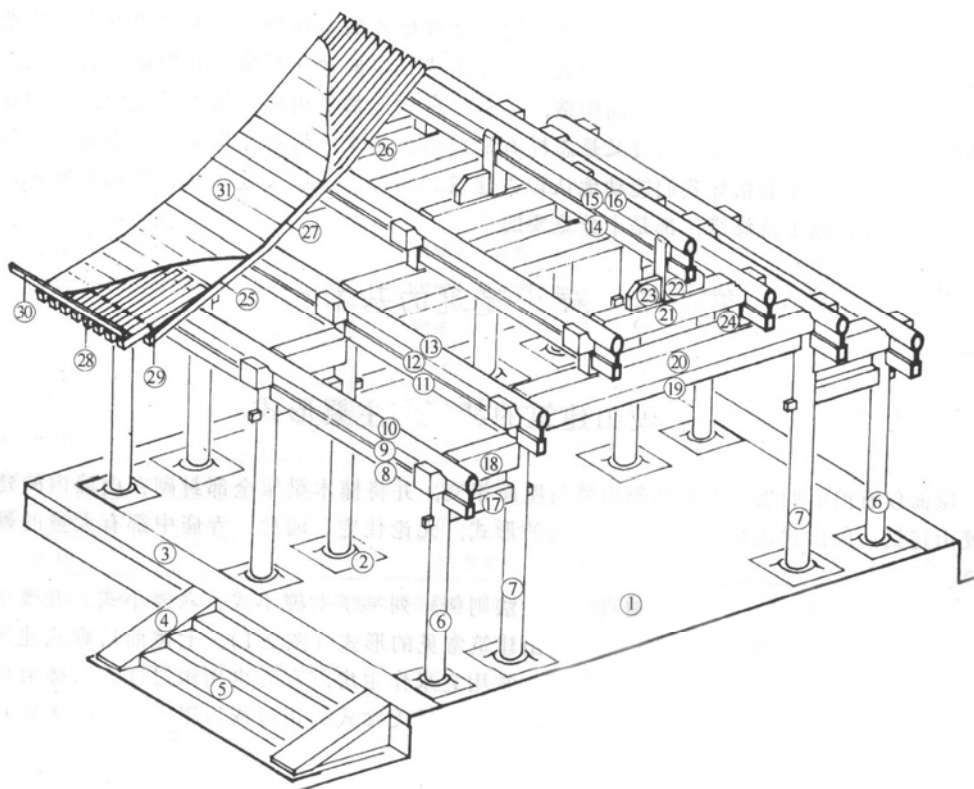
1) 整体思想

位置计算的基本目的是计算出一个构件中心点坐标值，对于有旋转角度的构件还需确定

其各方向上的旋转角度。古建筑领域对于一个建筑的描述,包括面宽和进深两个信息^[3],在此,笔者引入高方向这个描述,这样就可以完整描述一个构件在三维空间中的位置信息。

因而对于构件中心点坐标值的求解,需要确定三个方向上的坐标值,分别是面宽方向、进深方向、垂直方向,以下用三个变量来表示这三个方向上的坐标值, m (面宽), j (进深), g (高)。

(1) 首先计算面宽进深平面上的位置,直观地讲就是一个房子水平面上的坐标值。因为中国古典建筑在面宽、进深方向上以步架为单位构造的^[3],所以我们以步架为单位划分面宽、进深平面。硬山建筑的进深方向以间为单位划分,庑殿建筑的进深方向以间和步架为单位划分。这样二维平面可以分割出若干个进深缝和面宽缝,缝与缝的交点形成关键点,每个构件可以表示成连接某两个关键点的线,构件中心点的位置即可以由关键点的位置计算求出。如图2一个简单的三间七檩前后廊硬山建筑,其面宽进深平面可以步架为单位分割成如图4的情况,图2中10号构件檐檩连接了图3中A和B的两个关键点。图2中21号构件三架梁连接了图3中C和D两个关键点。



1. 台明 2. 柱顶石 3. 阶条 4. 垂带 5. 踏跺 6. 檐柱 7. 金柱 8. 檐枋 9. 檐垫板 10. 檐檩 11. 金枋
12. 金垫板 13. 金檩 14. 脊枋 15. 脊垫板 16. 脊檩 17. 穿插枋 18. 抱头梁 19. 随梁枋 20. 五架梁
21. 三架梁 22. 脊瓜柱 23. 脊角背 24. 金瓜柱 25. 檐椽 26. 脑椽 27. 花架椽 28. 飞椽
29. 小连檐 30. 大连檐 31. 望板

图2^[3] 硬山建筑木构架部位名称

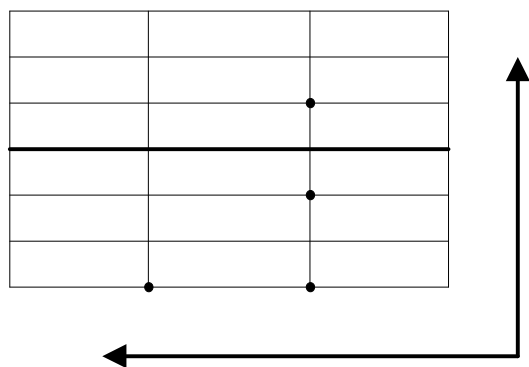


图 3 面宽进深方向切割图

垂直方向上的位置计算需要分两步：（1）按着垂直方向的基本单位分割，确定关键构件的垂直位置，主要是各层梁的垂直位置。如图 2 的建筑垂直方向可以分割成图 4 的情况



图 4 垂直方向切割图

（2）引入相关构件的描述。建筑就像积木一样是一层一层搭起来的，这样就很自然地有构件 A 在构件 B 之上，构件 B 在构件 A 之下的描述，则确定 A，B 中任一构件的垂直位置，另外一个构件的位置就可以求出。我们的做法是先求出关键构件，这里主要是指梁的位置，再通过相关构件的描述，求出其他构件的垂直位置。如图 3，13 号构件下金梁作为垂直方向上的关键构件将最先求得，20 号构件三架梁在 13 号构件之下，而 24 号构件瓜住又在 20 号构件之上。

本文所描述的构件位置计算均指无旋转角度的构件位置计算，关于有旋转角度的构件位置方向计算比较复杂，由于本文篇幅所限，在此不涉及。

2) 构件名称

大部分构件的位置是通过解析其名称而计算出的。本系统中用这样的三元组来描述构件名称

$$\text{Name}((m1,m2),(j1,j2),(g1,g2))$$

其中(m1,m2)表示进深方向上的跨度，(j1,j2)表示进深方向上的跨度，(g1,g2)表示高方向上的跨度。取房子门到后墙的方向为进深正方向，与进深方向夹角为 90 度的方向为面宽正方向，由下到上方向为高的正方向。按正方向将划分的缝依次标号如图 3、图 4 所示。

以这样的方式

图 3 标号为 6 的构件描述为檐住 ((2, 2), (7, 7), (1, 1))

标号为 13 的构件描述为檩 ((2, 3), (6, 6), (2, 2))

标号为 22 的构件描述为脊瓜柱 ((2, 2), (4, 4), (3, 4))

3) 位置计算规则设计

① 关键点位置计算

关键点位置计算分为面宽、进深、垂直三个方向上关键缝的位置计算，规定房子底面中心点为三维坐标原点，首先求得缝间距，则每缝的位置只需做一些简单的数学计算即可得，在此不加赘述。

a) 进深方向上各缝间距计算

进深方向上各缝间距即是指步架长度，古建领域，对于不同的制式，其步架规定不同，对于本系统所研究的三种制式，其步架规定如下表^[3]。

表 1 步架规则

制式	步架
硬山	4 * 柱径 ^[3]
庑殿	22 斗口 ^[3]
歇山	22 斗口 ^[3]

b) 面宽方向上各缝间距计算

计算面宽方向各缝的位置，首先需确定各缝间距，对于不同的制式采用不同的计算方法。
歇山

歇山面宽方向是以间为单位划分的，所以各缝间距即是各间面阔

明间面阔 = 柱高 * 1.25

次间面阔 = 柱高

梢间面阔 = 柱高

以图 3，3 间 7 檩硬山建筑为例，按图 4 所示面宽正方向描述，各缝间距依次为{梢间面阔、明间面阔、梢间面阔}

庑殿

庑殿面宽方向上的缝是以山面步架和各间面阔划分的，如图 5，一个 3 间 6 檩的庑殿其面宽方向上分割情况如图 7。

山面步架计算规则

庑殿山面步架计算采用庑殿推山法^[3]，以图 7 为例推山情况如下

x_1 =步架（见上一节步架计算规则）

$x_2 = x_1 - 0.1 x_1$

$x_3 = x_2 - 0.1 x_2$

....

算法如下

```
shanMianBuJia[1] = 正身步架
for I = 1->n
    shanMianBuJia[i] = shanMianBuJia[i-1] * 0.9
```


各间面阔计算规则

表 2 庀殿各间面阔规则

明间面阔	77 斗口
次间面阔	66 斗口
梢间面阔	55 斗口

则庀殿各缝间距依次为{（山面步架），（梢间面阔-山面步架和），次间面阔，明间面阔，次间面阔，（梢间面阔-山面步架和），（山面步架）}

竭山

竭山面宽方向上各缝的计算规则与庀殿基本相同，不同之处在于竭山山面步架仅有一缝，在此不加以赘述。

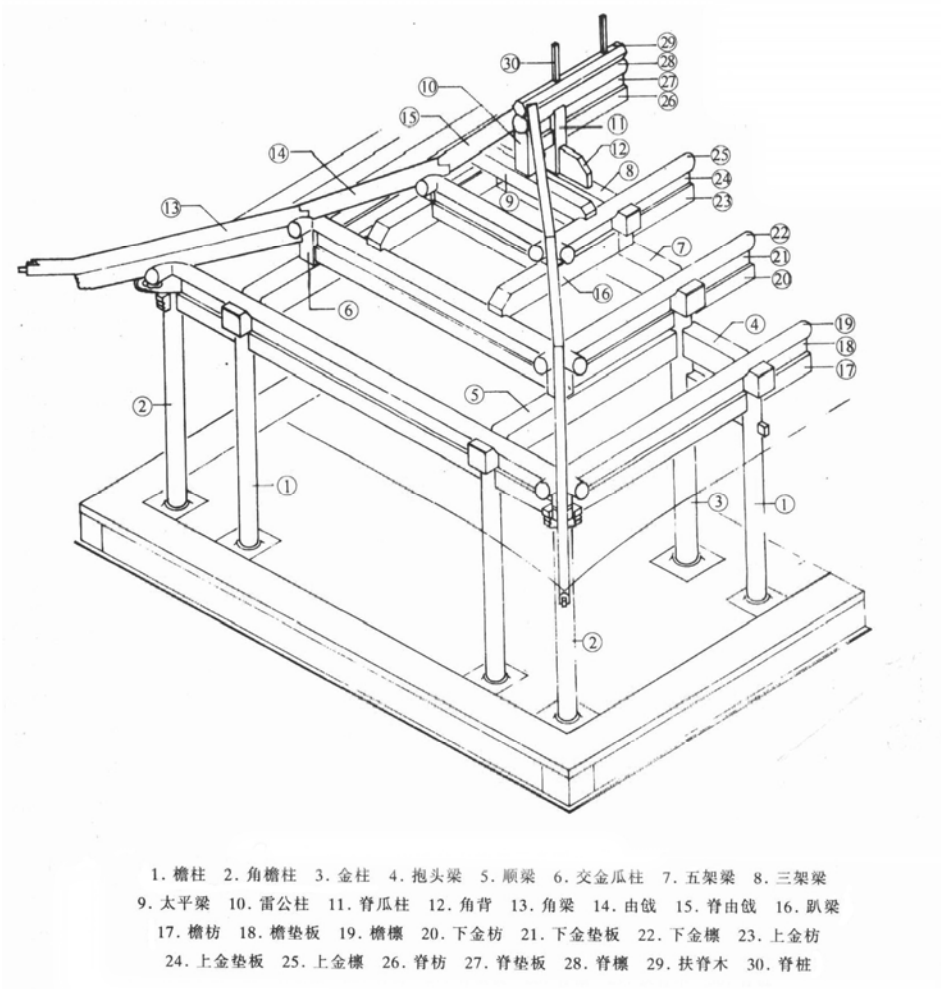


图 5 基本架构示意图^[3]

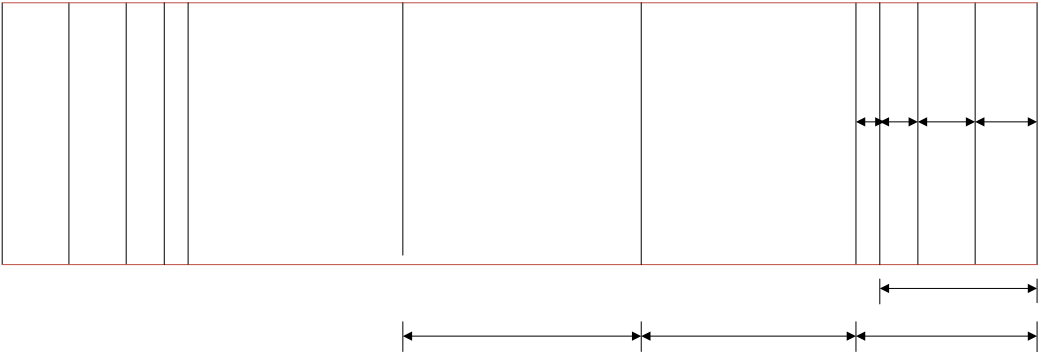


图 6

c) 垂直方向上各层位置计算

垂直方向上层的划分是檩为依据的，如图 3，檐檩位置为垂直第一层的位置，下金檩为第二层，上金檩为第三层，脊檩为第四层。

首先确定各层间距，在古建领域各层间距称为举架

举架 = 步架 * 举折^[3]（步架计算见进深方向上各缝计算）

不同形制的建筑举折有不同的规定。

硬山举折

对于一个 7 檩的硬山建筑，举折依次为：五举、七举、九举。我们归纳总结出如下的算法：

```
For i 1 -> n
  举折 = 0.5 + (i - 1) * 0.2
```

庑殿和竭山举折

庑殿和竭山举折有如下规定^[3]

表 3 庑殿竭山举折表

檩数	举折
5 檩	0.7, 0.9
7 檩	0.5, 0.7, 0.9
9 檩	0.5, 0.65, 0.75, 0.9
11 檩	0.5, 0.6, 0.65, 0.75, 0.9
13 檩	0.5, 0.6, 0.65, 0.75, 0.85, 0.9

其次确定第一层的绝对位置

```
IF 有斗拱
  第一层位置 = 柱高 + 斗拱高
Else
  第一层位置 = 柱高
```

最后，各层位置 = 前一层位置 + 间距，即可求出垂直方向上各层位置

② 构件面宽进深平面位置计算

由上面的计算，我们用以下三个数组分别描述面宽、进深、垂直各缝的位置，

mianKuan[1..n], jinShen[1..n], chuiZhi[1..n], 由 3.2 知这样描述这样构件 Name((m1,m2), (j1,j2),(g1,g2)), 另外用这样的三元组表示构件三个方向上的位置 {mPos, jPos, gPos}, 则构件面宽进深平面位置计算方法如下

```
If m1 = m2
    mPos = (mianKuan[m1] + mianKuan[m2])/2
else
    mPos = mianKuan[m1]
if j1=j2
    jPos = (jinShen[j1] + jinShen[j2])/2
else
    jPos = jinShen[j1]
```

③ 构件垂直方向位置计算

构件垂直方向位置计算方法如下

$$gPos = \text{相关构件.gPos} \pm (\text{构件高度} + \text{相关构件高度})/2$$

构件在相关构件之上, 则上式为+, 在相关构件之下, 则上式为-, 下表为相关构件表。

表 4 相关构件表

构件				相关构件				位置关系	说明
名称	面宽缝	进深缝	垂直层	名称	面宽缝	进深缝	垂直层		
柱础	m1,m2	j1,j2	g1,g2	柱	m1,m2	J1,j2	g1,g2	之下	同一层的 m,j,g 表示同一值, 下同
台基	m1,m2	j1,j2	g1,g2	柱础	m1,m2	J1,j2	g1,g2	之下	
檩垫板	m1,m2	j1,j2	g1,g2	檩	m1,m2	j1,j2	g1,g2	之下	
檩枋	m1,m2	j1,j2	g1,g2	檩垫板	m1,m2	j1,j2	g1,g2	之下	
梁	m1,m2	j1,j2	g1,g2	枋	#	#	g1,g2	之上	#表任意值, 下同
瓜柱	m1,m2	j1,j2	g1,g2	梁	#	#	(g1-1),(g2-1)	之上	
随梁枋	m1,m2	j1,j2	g1,g2	梁	#	#	1, 1	之下	
抱头梁	m1,m2	j1,j2	g1,g2	檩枋	#	#	1, 1	之下	
穿插枋	m1,m2	j1,j2	g1,g2	檩枋	#	#	1, 1	之上	
趴梁	m1,m2	j1,j2	g1,g2	垫板	#	#	g1,g2	之下	$g1 \leq 2 \ \&\& \ g2 \leq 2$
				枋	#	#	g1,g2	之下	$g1 > 2 \ \&\& \ g2 > 2$
交金瓜柱	m1,m2	j1,j2	g1,g2	趴梁	#	#	g1,g2	之上	
太平梁	m1,m2	j1,j2	g1,g2	垫板	#	#	$g1-1, g2-1$	之上	
雷公柱	m1,m2	j1,j2	g1,g2	太平梁	#	#	g1,g2	之上	
平板枋	m1,m2	j1,j2	g1,g2	檐柱	#	#	g1,g2	之上	
大额枋	m1,m2	j1,j2	g1,g2	平板枋	#	#	#	之下	
由额垫板	m1,m2	j1,j2	g1,g2	大额枋	#	#	#	之下	
小额枋	m1,m2	j1,j2	g1,g2	由额垫板	#	#	#	之下	
踩步金	m1,m2	j1,j2	g1,g2	柱	m1,m2	J1,j2	g1,g2	之下	同一层的 m,j,g 表示同一值
踩步金枋	m1,m2	j1,j2	g1,g2	柱础	m1,m2	J1,j2	g1,g2	之下	

4) 实验结果

笔者将本文的设计方法付诸于实现，并嵌入到古建动画自动生成系统中，成功实时生成了多种类型古建筑三维场景，实现的古建筑种类涵盖了现存的几乎所有明清建筑制式，主要包括以下几种。

表 5 系统可生成古建类型

形制	规格
庑殿	5 7 9 11 间 7 9 11 13 檩
竭山	5 7 9 11 间 7 9 11 13 檩
硬山	3 5 7 9 11 间 5 7 9 11 13 檩

下面是由本系统实时生成的 7 间 9 檩竭山、庑殿、硬山建筑三维模型。

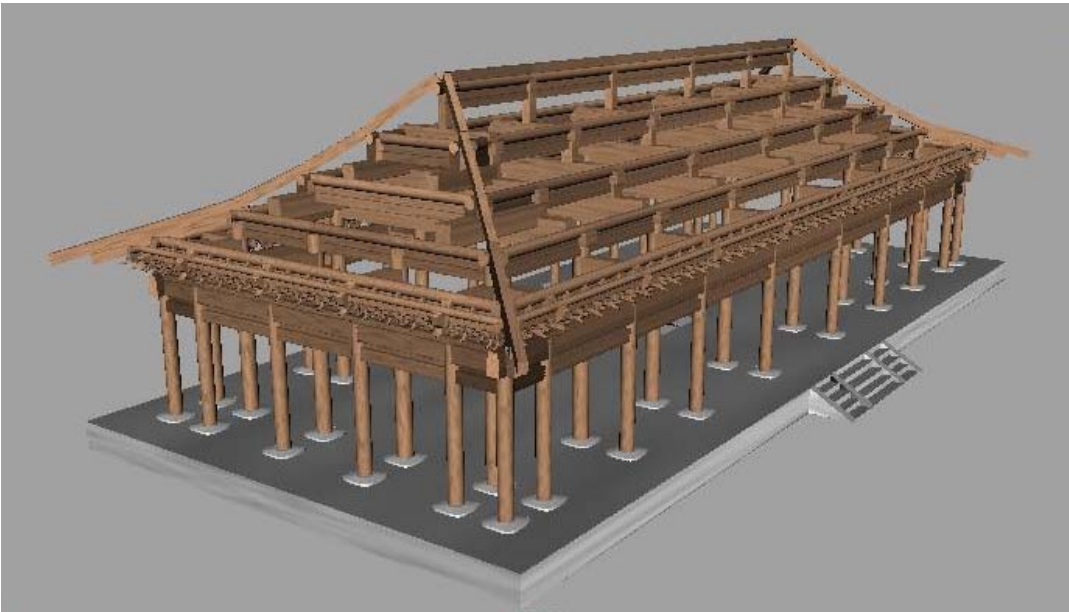


图 7 7 间 9 檩庑殿建筑生成图

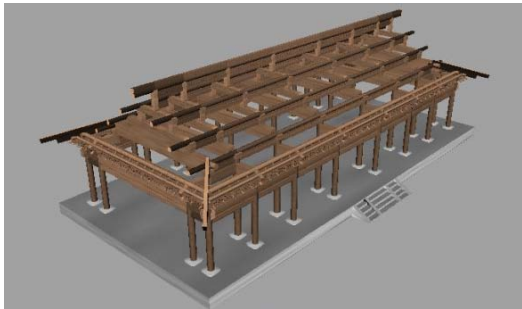


图 8 7 间 9 檩竭山建筑生成图

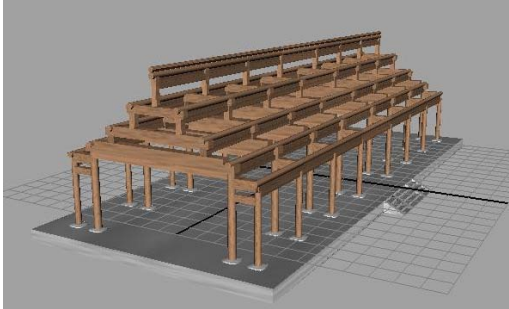


图 9 7 间 9 檩竭山建筑生成图

4 总结

中国古典建筑建筑过程动画自动生成系统对古建筑知识普及具有重要意义, 但本系统还存在许多不足之处, 如对构件位置计算是用代码实现的, 在以后的工作中考虑用推理机实现。另外在研究过程中, 由于缺少一些古建规则, 本系统的一些参数是一部分经验总结的结果, 还有一小部分存在猜测的问题, 希望随着古建筑领域研究的发展进一步完善。

5 感谢

感谢项目组的孙凯和王巍锋同学, 他们完成了构件顺序分析工作, 感谢北京房修二公司的倪原总工程师为我们提供了极有价值的古建知识, 感谢项目组的各位老师。

参考文献

- [1] 金小刚, 鲍虎军, 彭群生. 计算机动画技术综述. 软件学报, 1997, 8(4): 241.
- [2] 陆汝钤, 张松懋. 从故事到动画片——全过程计算机辅助动画自动生成. 自动化学报. 2002, 28(3): 321.
- [3] 马炳坚. 中国古建筑木作营造技术 (第二版). 北京: 科学出版社, 2003.
- [4] 梁思成. 清式营造则例. 北京: 清华大学出版社, 2006.
- [5] 米德, 阿里马. Maya 完全学习手册. 北京: 清华大学出版社, 2005.

作者简介

尹梅芳, 女, 北京工业大学在读硕士研究生; 朱宫瑾, 女, 北京工业大学在读硕士研究生

IT应用与经管教学融合的探讨

张 耀¹ 崔锦荣²

(1.大连东软信息学院, 大连, 116023; 2.大连大学经济管理学院, 大连, 116622)

摘 要: IT 应用与经管教学相融合是培养经管人才的重要方式和必然趋势。这种教学模式能够使教师的教学效率更高、效果更佳; 又使学生实际操作能力大为增强, 为将来工作奠定良好基础。这种模式的构建, 要以“能力”培养为中心, 以多媒体和 Internet 网络为教学媒介, 在完善师资知识技能结构的基础上实施教学。

关键词: IT; 经管教学; 探讨

An Exploration of IT Application to Business Administration Education

ZHANG Yao¹ CUI Jin-rong²

(1. Dalian Neusoft Institue of Information, Dalian, Cihina, 116023;

2. School of Economic Management, Dalian University, Dalian, Cihina, 116622)

Abstract: The integration of information technology into business education is applied in the current and developing business courses to cultivate business professionals. This model makes teaching more efficient and improves students' ability to solve real-world problems so as to lay a solid foundation for their future career. Construction of this model should be as follows: focus on the development of students' problem solving ability, employ multimedia and Internet as the teaching medium and advance teachers' knowledge and skills.

Keywords: IT; Business Administration Education; Exploration

在社会经济发展日新月异的今天, 在信息化浪潮的冲击下, IT 技术应用越来越普及, 已经渗透到各个领域, 社会和企业对经济管理人才的知识技能也有了不同以往的更高要求。如何培养适应信息化和企业管理现代化人才, 已经成为经管教育必须直面的一个现实问题。

1 IT应用与经管教学融合的必要性

1) IT 软件应用是经管人才知识技能结构不可或缺的组成部分。当前, 国民经济步入了一个结构调整、产业升级的新阶段, 需要大批掌握现代管理理论和方法、并同时具备 IT 应用技能的人才。不难想象, 一个不懂 IT 应用的经管人才, 在今天是否还能适应现代企业管理的要求。所以, 经管人才的知识技能结构必须调整充实, 将 IT 应用技能融入其中, 使之趋于完善。

2) 顺应信息化潮流是经管教学的必然趋势。在社会和企业需要大批具备 IT 应用技能人才的环境下, 经管学科教育不仅仅是一个适应社会对专业和职业需求的问题, 更重要的是一个适应高等教育环境变化、面向全球化、应对信息化的转变过程。由于社会经济的飞速发展, 知识更新周期缩短且频率加快, 经管教学跟不上社会发展的节奏, 其教学模式和方法也已经严重滞后。因此, 将 IT 应用于经管教学, 融二者于一体的教学模式, 是经管人才培养路径的一个现实选择。

从表 1 中可看到, 在国民经济各领域中, 具有 IT 应用技能的经管业务岗位几乎覆盖了社会和经济的所有部门, 呈现了较为广阔的职业发展前景。

表 1 国民经济各部门中的经济管理预期职业岗位

	职业领域	举例	岗位群
1	政府综合协调部门	发改委、财政局、国有资产局、劳动局等	地区和行业宏观统计和财务系统的操作员
2	政法和社会保障部门	公安局、司法局、安全局、民政局等	社会保障各类统计和财务的操作员
3	社会公共事业管理部门	教育苟、文化局、广电局、卫生局、体育局等	行业统计和财务的操作员
4	国民经济各个专业管理部门	交通局、农业局、商业局、城乡建设局等	行业统计和财务的操作员
5	监督检查部门	审计局、工商局、税务局、环保局、物价局、统计局、标准计量局等	财税统计和财务的操作员
6	政党组织、社会团体、人民解放军各级部门	从事统计、党费团费管理	人事统计和财务的操作员
7	全国各企事业单位, 科研院所	化工、建材、冶金、机械、纺织、轻工、电子、医药等企业	产供销物流统计和财务的操作员
8	医院、卫生院及医疗机构	门诊、住院、药房信息管理	统计、财务、病人、药品等信息管理的操作员
9	商场、超市、批发机构等	开票、收银、库存、采购、销售等	库存、统计和财务的操作员
10	档案馆、图书馆、情报检索机构、信息中介	数据库建设、信息检索服务	信息录入、检索与服务的操作员
11	饭店、宾馆、旅行社等	点菜、客房、组团、接团、结帐收银	统计、财务及旅游业务信息管理的操作员
12	学校教务管理	学籍、教材、课程、师资、实验实训、固定资产等管理	统计、财务及教务信息管理的操作员
13	金融、保险、税务、审计机构	财产险、人寿险、银行信贷、税收监控、财产审计等	统计分析、综合管理的操作员
14	IT 软件服务业	软件公司、咨询公司、网站等	信息采编员、操作员等

2 经管人才培养面临的问题

1) 经管人才知识技能的结构性错位。信息化社会需要的是信息化人才, 企业也一直对这类人才青睐有加, 但现实情况却不容乐观, 经管人才知识结构单一, 缺乏 IT 应用技能, 许多大学毕业生与社会期望和企业需求不相匹配, 出现人才结构性错位, 也造成人力资源浪费。

2) 教师知识技能结构缺陷的制约。在经管教学中运用 IT 技术, 利用多媒体和 Internet 手

段教学,不但对教师的经济管理基本理论有很高的要求,而且对IT技术的应用能力也有较高的水准要求。但遗憾的是,兼具二者能力的教师还不很多,因而对教学的影响很大。

3) 学用脱节的教学模式。经济管理是实践性很强的学科,以往纸上谈兵式的黑板加粉笔、理论加书本的教学,严重脱离社会实际和企业现实,难以满足社会信息化和企业管理现代化的要求。

3 经管教学与IT应用相融合模式的构建

1) 以能力培养为目标的教学理念

① 教学模式上的探索。改变过去那种重理论轻能力的教学理念,按照社会和企业信息化对人才的要求,调整和修订教学大纲。在教学手段上,充分利用现代科技手段,融IT应用于教学实践;在内容的把握上,教学重心由基础理论的讲授转向实际能力的培养。最终结果是,教师实现教学目标,学生达到学习目的,社会和企业获得合格的人才。

② 教学形态上的变化。传统的教学方法与手段已经适应不了社会经济发展形势的新要求,经管教学在IT技术应用和Internet网络的支撑下,变传统静态教学为适应环境要求的动态教学,以不断更新、紧扣社会现实的教学内容引导和激发学生的兴趣,即可以丰富课堂教学内容,还能够调动学生的积极性。

③ 时空延伸全天候式学习。借助于IT技术和Internet,教学可以跨越课堂上下和教室内外,保证学习的系统性和连续性,增加了学习、锻炼和实践的机会,同时促进“第一课堂”和“第二课堂”的有机结合,对延伸和扩展知识,培养和演练学生解决问题的实际能力均具有重要作用。

2) 理论与实践相结合教学体系

基于经管教学目标,在掌握现代经济管理基本原理和方法的基础上,使学生能够熟练运用相应的经管软件,初步具备解决现实经济管理问题的能力,形成“理论教学与实践教学相互交叉、理论学习与动手能力培养相互促进、知识掌握与创新能力培养有机协调的良性循环体系。”

① 基础性的理论教学。理论是基础,理论指导实践,是实践的灵魂。管理问题具有很强的现实背景,在学习课程内容时,学生常常感觉抽象空洞,遇到具体问题时无从下手,不知所以然。在教学过程中,按照务实基础、拓宽口径、增强能力、突出应用、提高素质的原则,对原有的理论课程体系及教学内容进行调整,明确理论学习的意义,构建与能力培养相适应的理论教学体系。

② “能力性”的实践教学。现代经管教学,在很大程度上既强调对理论的掌握、分析、理解,又强调理论与实际紧密结合。实践教学在经管人才能力培养体系中起着关键性的作用,它是提高学生动手能力的主要途径。根据经济管理学科内容多、综合性强的特点,采取多种形式,构建从课堂内系统的综合性实践技能训练、到课外的自助式开放实践和校外实训相结合的实践教学体系。实践环节可以帮助学生沟通很多专业课程的联系,融会贯通地加以理解,改变了各门课程相互脱节的状况。实践教学既是理论教学的延续,同时又是检验和巩固理论教学效果的重要环节,二者相互促进,相得益彰。

3) 充分利用IT应用软件和多媒体及Internet网络

运用IT应用软件和多媒体教学,可以为学生提供全方位、多渠道、最直接的视听感受,

拓展教学空间,提高教学效率。

① 基于 IT 技术的多媒体教学。在教学方式上,通过多媒体为学生获取更多的知识和信息创造条件,以多媒体电子课件为基础,充分利用声光电的媒介效果,再辅之以教学录像,形象化生动化的教学成为经管教学的主旋律。

② 整合 Internet 资源,进行 Online 教学。Internet 上有很多内容丰富、形式多样的资料,有最新的企业案例、在线视频、网上操作软件,这些不但是宝库的教学资源,而且是经管教学不可多得的实践平台。把握经济管理领域最新的国内外动态信息,将最新的网上案例直接引入课堂,这种课堂与社会直接对接的教学方式,大大缩短了学生获取知识的周期。现在的学生生长、学习和生活在信息发达的时代,视野开阔,见多识广,实施 Online 教学,则更能与学生产生共鸣,使教与学融为一体。

③ IT 应用软件的操作教学。限于客观条件,不可能每堂课每个知识点都去实习,最现实的办法就是在虚拟环境下进行模拟操作,以增强实践效果。IT 应用软件就是虚拟环境下的仿真操作平台,通过这个平台,及时消化理解课堂内容,体验真实感受,提高学习效率。

4) 强化师资队伍建设

师资是教学的重要基础条件,高水平的师资乃是培养高素质经管人才的关键要素。①优化师资队伍的知识结构。通过多种形式培训师资,补充 IT 技能,完善知识结构,以尽快达到教学要求。②走出校门丰富实践经验。通过制度性的激励措施,引导鼓励教师走出校门,深入工商企业,进行学习调研或直接参与企业的经营管理、项目开发等工作。这样,不但能够增强教师自身的实践能力、丰富实践经验,而且能够及时了解企业实际,引进实际案例,保持课堂内容与企业实际的联动互补。

4 结语

IT 技术在经管教学中的应用,基于这样两个层面:一是在教学方面,教师借助 IT 的技术手段辅助教学,从而提高教师教学效果和学生的学习效率。在教学过程中及时把握学科发展的趋势、跟踪学科发展前沿的最新成果,充实进课堂。二是在学习方面,在理解经济管理基本理论的基础上,学生借助于管理软件围绕企业管理的实务问题,掌握常用管理软件的基本应用,为将来奠定良好的工作基础。

参考文献

- [1] 李利平.对高校计算机公共基础教学的探讨[J].计算机教育,2009(10):63-65.
- [2] 杜文洁,刘春颖.对多媒体教学及其改革的思考[J].计算机教育 2009(6):51-53.
- [3] 陈伊立.面向目标的经管类学生计算机深入教学创新[J].2009(3):59-64.
- [4] 刘垣.应用型本科计算机公共基础教育改革刍议[J].计算机教育,2008(12):15-17.

作者简介

1. 张耀,男,山西稷山人,硕士研究生,1958年11月生,副教授,研究领域:经济管理和高等教育。
2. 崔锦荣,女,1960年8月生,副教授,研究领域:会计学和财务管理。

第 2 部分

数字信号处理

Short-Term Load Forecasting Based on Dynamic Recurrent Fuzzy Neural Network

GE Chao ZHANG Jing-chun SUN Yan-bin SUN Li-ying

(Hebei Polytechnic University College of Information, Hebei Tangshan 063009)

Abstract: Because of the limitations of general fuzzy neural network BP algorithm, a novel dynamic recurrent fuzzy neural network, which is applied to modeling problems in electric power system short-term load forecast, is proposed. The fuzzy inference function is realized easily by using a product operation in the network. Introducing local recurrent units to hidden layer, the proposed method can overcome the limit of BP algorithm. The performance of the proposed model is evaluated based on a North China power grid operational load data. Simulation results showed that the forecasting precision of the dynamic recurrent fuzzy neural network was enhanced by 4.4%, while the time cost was only increased by 2.1 seconds, indicating that the improved algorithm could gain better forecasting effect.

Keywords: Short-Term Load; Dynamic Recurrent; Fuzzy Neural Network; Forecast Model

1 Introduction

Short-term load forecast is of great significance for the economic, reliable and safe operation in electric power systems. With the market access of electric industry, there is a higher demand for the precision of short-term load forecast. Now there are two ways for short-term load forecast^[1-3]: Time series technique and artificial intelligence (based on neural network). Time series model, which is easily to comprehend and operate, is unable to cope with multivariate problems or heteroscedasticity, so the forecasting precision in actual application is limited to a large extent. With the development of artificial intelligence, neural network technique is widely used, esp. in the forecasting of more complicated, non-linear systems. Fuzzy neural network, which is able to cope with more influential factors than traditional neural network, has achieved a better effect in actual forecasting practice. However, most fuzzy neural networks are based on BP analysis, so there is a structure defect^[4-5] that it may come into a state of partial extreme value.

A novel dynamic recurrent fuzzy neural network (DRFNN) is proposed based on the characteristics of short term load forecast. Multiplication Operation is used in rule layer to ensure the activation degree $\omega_j(k) \leq 1$ for each rule, thus easier to realize the fuzzy inference function. Through introducing delay unit into rule layer, static networks have dynamic features; Network activation degree at K times for each rule not only refers to the activation value after computing, but

also includes all the activation value contribution at previous times. The matlab simulation program, which is designed with this model, is applied to the changeable and complex short term load forecasting problems. Taking the electric network data of a north China city as an example, a short term load forecasting is conducted. The result shows that this technique is of good forecasting precision degree and high reliability.

2 Dynamic Recurrent Fuzzy Neural Networks (DRFNN)

As figure 1 show, network topology structure has 5 layers, namely input layer, fuzzy layer, rule layer, and output layer. There is recursive layer neurons in rule layer, and the neurons are recursive as a result of internal feedback connections. So dynamic response is acquired that can simplify the network model.

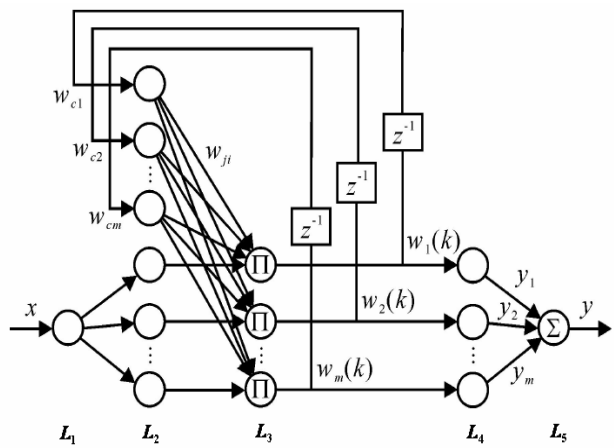


Fig.1 Frame of dynamic recurrent fuzzy neural network

L_1 is input layer, directly connected to input vector, transmits the input value to next layer. L_2 is fuzzy layer, each node represents a language variable value, such as S_2, S_1, C_E, B_1, B_2 and so on. It is used to compute membership function (The input value belongs to fuzzy set of various language variables). According to the real situation, Gaussian function is chosen as the input membership function^[6-7]:

$$\mu_j = \exp \left[- \left(\frac{x(k) - x_j}{\delta_j} \right)^2 \right] \tag{1}$$

Where x_j is the center of the j subjection function, δ_j is the width of the j subjection function, $x(k)$ is the input of the time of k , $j = 1, 2, L, m$, m is the fuzzy divisions.

L_3 is rule layer, each node, which represents a fuzzy rule, has a function of matching first component of fuzzy rules., and then figure out activation degree $\omega_j(k)$ of each rule. When the recursive layer is introduced, according to the network shown in figure 1,

$$\omega_j(k) = \left[\prod_{i=1}^m \omega_{ci}(k) W_{ji} \right] \exp \left[- \left(\frac{x(k) - x_j}{\delta_j} \right)^2 \right] \quad (2)$$

Where W_{ji} is the connection weight from the j node of recursion layer to the i node of the rule layer, $\omega_{ci}(k)$ is the activity of the i node of the rule layer,

$$\omega_{ci}(k) = \omega_i(k-1), i=1, 2, L, m \quad (3)$$

The formula (2) and (3) show that the veracity of the network identification is increased and the static network has the dynamic characteristic. Because the recursion layer is introduced in the network structure, the activity $\omega_j(k)$ of each rule in the network at the time of k not only includes the activity value μ_j which is obtained from the current input, but also includes the contribution of the former activity value $\prod_{i=1}^m \omega_i(k-1)$.

Different from other dynamic recursive networks, this paper puts forward multiplication operation in rule layer. In this way, not only is the activation degree for each rule ensured ($0 < \omega_j(k) \leq 1$), but also the fuzzy reasoning function is easier to achieve, because the fuzzy rule mode is easy for operators to understand.

L4 is the defuzzification layer, which realized the defuzzification operation. The definition of each node activity in this layer is:

$$\phi = 1 / \sum_{i=1}^m \omega_i(k) \quad (4)$$

The output of the i node in this layer is:

$$\overline{\omega_i}(k) = \phi \omega_i(k) \quad (5)$$

The fifth layer is the output layer. The output of the network at the moment k is:

$$y(k) = \sum_{i=1}^m \mathcal{Y}_i(k) \overline{\omega_i}(k) = \sum_{i=1}^m \mathcal{Y}_i(k) \omega_i(k) / \sum_{i=1}^m \omega_i(k) \quad (6)$$

Where $\mathcal{Y}_i(k)$ is the conclusion of the i fuzzy rule at the moment k .

3 DRFNN Learning Algorithms

When the fuzzy division is confirmed, the network parameters need to be trained. They are the connection weight W_{ji} from recursion layer to rule layer, the center x_j and width δ_j of the input subjection function, the connection weight \tilde{y}_i from hidden layer to output layer.

The objective function can be defined by Mean Square Error (MSE):

$$MSE = \frac{1}{N} \sum_{k=1}^N E_p(k) \quad (7)$$

Where N is the sample number, $E_p(k)$ is the instantaneous square error:

$$E_p(k) = \|y_d(k) - y(k)\|^2 \quad (8)$$

Where $y_d(k)$ is the true sample output at the moment t , $y(k)$ is the output of DRFNN at the moment t .

From formula (2) and (3), we can find:

$$\omega_{cj}(k) = \omega_j(k-1) = \left[\prod_{i=1}^m \omega_{ci}(k-1) W_{ji} \right] \times \exp \left[- \left(\frac{x(k-1) - x_j}{\delta_j} \right)^2 \right] \quad (9)$$

The formula (9) implies that $\omega_{cj}(k)$ is the dynamic recurrence course which is depended on the foretime connection weight. So the relevant BP algorithm is the dynamic BP algorithm^[8-9].

The parameters are adjusted by the gradient descent method, which can find the minimum parameter for the object function E_p . Learning algorithm for adjusting specific parameters is as follows:

3.1 Connect Weight (W_{ji}) Adjustment from Recursive Layer to Rule Layer

$$W_{ji}(k+1) = W_{ji}(k) - \eta_{\omega} \partial E_p / \partial W_{ji} |_k \quad (10)$$

Where η_{ω} is the learning rate coefficient. The big learning rate coefficient can make the algorithm converge fast, but prone to oscillation. The small learning rate coefficient can slow oscillation, but the convergence gets slower, and then the principle of their choice must accord to the actual identification. We can get as follow by using chain rule:

$$\frac{\partial E_p}{\partial W_{ji}} = \frac{\partial E_p}{\partial y(k)} \frac{\partial y(k)}{\partial \omega_j(k)} \frac{\partial \omega_j(k)}{\partial W_{ji}} \quad (11)$$

We can get the derivative by formula (8) :

$$\partial E_p / \partial y(k) = -(y_d(k) - y(k)) \quad (12)$$

by formula (6):

$$\frac{\partial y(k)}{\partial \omega_j(k)} = \frac{y_j(k) - y(k)}{\sum_{i=1}^m \omega_i(k)} \quad (13)$$

by formula (2) and (9):

$$\frac{\partial \omega_j(k)}{\partial W_{ji}} = \left[\prod_{n=1}^m \omega_n(k-1) W_{jn} \right] \exp \left[- \left(\frac{x(k) - x_j}{\delta_j} \right)^2 \right] \times \left(\frac{1}{W_{ji}} + \frac{1}{\omega_j(k-1)} \frac{\partial \omega_j(k-1)}{\partial W_{ji}} \right) \quad (14)$$

This will constitute the gradient $\partial \omega_j(k) / \partial W_{ji}$ dynamic recurrence relation, and this is similar with the time back-propagation learning algorithm.

3.2 Input Value Membership Function Centre(x_j)Adjustment

$$x_j(k+1) = x_j(k) - \eta_x \partial E_p / \partial x_j |_k \quad (15)$$

Where η_x is the learning rate coefficient, the principle of selection is introduced as before. According to chain rule we can get the derived function as follow:

$$\frac{\partial E_p}{\partial x_j} = \frac{\partial E_p}{\partial y(k)} \frac{\partial y(k)}{\partial \omega_j(k)} \frac{\partial \omega_j(k)}{\partial x_j} \quad (16)$$

By equation (2) and (9) we can get as follow:

$$\frac{\partial \omega_j(k)}{\partial x_j} = \left[\prod_{i=1}^m \omega_i(k-1) W_{ji} \right] \exp \left[- \left(\frac{x(k) - x_j}{\delta_j} \right)^2 \right] \times \left[2 \left(\frac{x(k) - x_j}{\delta_j} \right) + \frac{1}{\omega_j(k-1)} \frac{\partial \omega_j(k-1)}{\partial x_j} \right] \quad (17)$$

The equation (17) will constitute the gradient $\partial \omega_j(k) / \partial x_j$ dynamic recurrence relation. The equation (16) can be get by the equation (12), (13) and (17).

3.3 Input Value Membership Function Width (δ_j) Adjustment

$$\delta_j(k+1) = \delta_j(k) - \eta_\delta \partial E_p / \partial \delta_j|_k \quad (18)$$

Where η_δ is the learning rate coefficient. According to chain rule we can get the derived function as follow:

$$\frac{\partial E_p}{\partial \delta_j} = \frac{\partial E_p}{\partial y(k)} \frac{\partial y(k)}{\partial \omega_j(k)} \frac{\partial \omega_j(k)}{\partial \delta_j} \quad (19)$$

$$\frac{\partial \omega_j(k)}{\partial \delta_j} = \left[\prod_{i=1}^m \omega_i(k-1) W_{ji} \right] \exp \left[- \left(\frac{x(k) - x_j}{\delta_j} \right)^2 \right] \times \left[2 \frac{(x(k) - x_j)^2}{\delta_j^3} + \frac{1}{\omega_j(k-1)} \frac{\partial \omega_j(k-1)}{\partial \delta_j} \right] \quad (20)$$

The equation (20) will constitute the gradient $\partial \omega_j(k) / \partial \delta_j$ dynamic recurrence relation. The equation (19) can be get by the equation (12), (13) and (20).

3.4 Connect Weight \tilde{y}_i Adjustment from Hide Layer to Output Layer

$$\tilde{y}_i(k+1) = \tilde{y}_i(k) - \eta_{\tilde{y}} \partial E_p / \partial \tilde{y}_i|_k \quad (21)$$

Where $\eta_{\tilde{y}}$ is the learning rate coefficient. According to chain rule we can get the derived function as follow:

$$\frac{\partial E_p}{\partial \tilde{y}_i} = \frac{\partial E_p}{\partial y(k)} \frac{\partial y(k)}{\partial \tilde{y}_i(k)} \quad (22)$$

By equation (6) we can get as follow:

$$\frac{\partial y(k)}{\partial \tilde{y}_i(k)} = \omega_i(k) / \sum_{i=1}^m \omega_i(k) \quad (23)$$

By equation (12) and (23) we can get equation (22).

4 Dynamic Recurrent Fuzzy Neural Network of Short Term Load Forecast

In order to verify the validity of DRFNN, a computer simulation about DRFNN and traditional FNN is conducted, taking the electric power system data from 1st May to 15th June in 2008 of a North China city as samples. The random number (W_{ji} 和 \tilde{y}_i is $[0, 1]$ respectively) is chosen. The

initial input membership function is Gaussian function. The end condition is as follows: 2000 times circulation or the error reaches 0.001.

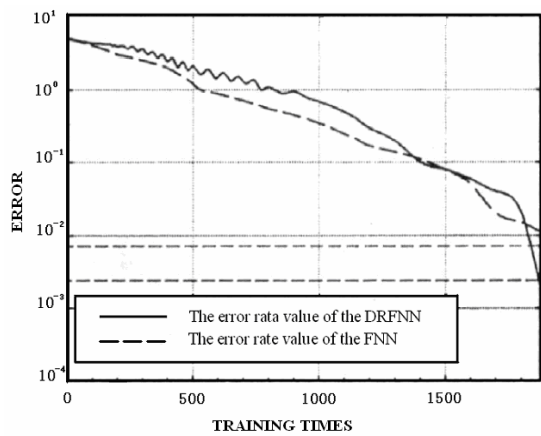


Fig.2 Training error comparison of DRFNN and FNN

As figure 2 and 3 shows, according to the comparison between DRFNN and FNN training error curve chart, the network gains dynamic feature after introducing recursive delay unit. Therefore, the learning ability of neuron network model is improved, and the lower training error is reached.

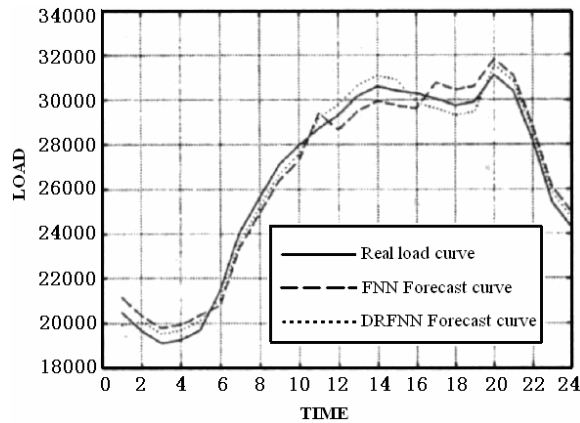


Fig.3 Prediction comparison of DRFNN and FNN

As figure 3 shows, DRFNN forecast curve is more similar to the real load curve than FNN. It is proved that FNN forecast is improved by multiplication operation introduced into recursive layer to some extent. Figure 1 shows that the forecast time of the two models are approximately the same. It is proved that DRFNN forecast model ensures forecast efficiency while increasing the forecast accuracy.

Tab.1 Prediction comparison of DRFNN and FNN

Forecastmodel	Forecast accuracy/%	Time/s
DRFNN	97.9	75.9
FNN	93.5	73.8

5 Conclusions

The short term load forecasting model based on DRFNN in this article can improve the learning ability about the fuzzy neural network effectively. Through Simulation Verification, it is proved that both the convergence rate and forecasting accuracy of DRFNN is increased to some extent compared to traditional FNN, avoiding the problem of partial optimum. This network, which has been proved effective through illustrative examples, can be applied to long term forecast in electric power system after the research methodologies are improved and perfected.

References

- [1] 张友旺. 基于动态递归模糊神经网络的动态系统辨识[J]. 中南工业大学学报, 2003, 34(3): 277-280. ZHANG You-wang. Identification of dynamic system based on dynamic fuzzy neural network[J]. Journal of Center South University Technology, 2003, 34(3): 277-280.
- [2] 高山, 单渊达. 基于径向基函数网络的短期负荷预测. 电力系统自动化, 1999, 23(5): 31-34. GAO Shan, SHAN Yuan-da. A new neural network short-term load forecasting algorithm using radial basis function network. Automation of Electric Power Systems, 1999, 23(5): 31-34.
- [3] C. J. Lin, C. C. Chin, Prediction and identification using wavelet-based recurrent fuzzy neural networks, IEEE Trans. Fuzzy Systems 34(5) (2004) 2144-2154.
- [4] W. Yu, State-space recurrent fuzzy neural networks for nonlinear system identification, Neural Process. Lett. 22(3) (2005)391-404.
- [5] 赵登福, 张涛, 杨增辉, 等. 基于 GN BFGS 算法的 RBF 神经网络短期负荷预测. 电力系统自动化, 2003, 27(4): 23-27. ZHAO Deng-fu, ZHANG Tao, YANG Zeng-hui, et al. Short-term load forecasting using radial basis function(RBF) neural networks based on GN-BFGS algorithm. Automation of Electric Power Systems, 2003, 27(4): 23-27.
- [6] C. F. Juang, ATSK-type recurrent fuzzy network for dynamic systems processing by neural network and genetic algorithms, IEEE Trans. Fuzzy Systems. 10(2) (2002)155-170.
- [7] Lee Ching-hung, Teng Ching-cheng. Identification and control of dynamic systems using recurrent fuzzy neural networks[J]. IEEE Trans on Fuzzy Systems, 2004, 8(4): 349-366.
- [8] PAI P F, HONG W C. Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms[J]. Electric Power Systems Research, 2005, 74(3): 417-425.
- [9] 姜勇. 电力系统短期负荷预测的模糊神经网络方法[J]. 继电器, 2002, 3(7): 11-13. Jiang Yong. Fuzzy neural network for short-term load forecasting[J]. Relay, 2002, 3(7): 11-13.

Author

Ge Chao(1980.9—), male. Now is a teacher of Hebei Polytecnic University. His research area is the application of artificial intelligence in the electric power system. He can be reached by E-mail: gechao365@heut.edu.cn

Address: College of Information, Hebei Polytecnic University, No. 46 Xinhuaixi Road, Tangshan, Hebei Province, Postcode: 063009

基于信息熵的模糊聚类信誉评价体系

宫尚宝 郭玉翠

(北京邮电大学理学院 北京 100876)

摘 要: 在诸如文件共享等无中心的 Peer-to-Peer 网络中, 对等节点具有匿名性和高度自治性的特点, 由于缺乏对与之交互节点的可信程度的知识, 在交互过程中节点可能会遭遇恶意推荐和欺骗等威胁。为了消除这类威胁, 本文提出了一种基于信息熵的模糊聚类信誉评价体系, 利用信息熵来确定最优的聚类数目, 利用概率密度函数来选择初始聚类中心, 引入余弦相似函数作为相似程度。通过与已知的 EigenRep 算法比较, 仿真结果证明了新的评价体系能够很好地消除网络节点的恶意推荐和欺骗。
关键词: 模糊聚类; 信任; 信息熵; 相似度

The Credibility Evaluation System Based on Fuzzy Clustering Of the Entropy

GONG Shang-bao GUO Yu-cui

Abstract: The risk involved with the transactions for without prior knowledge about each other's reputation has to been managed in decentralized peer-to-peer file-sharing networks, due to the anonymous and self-organization nature of peers. In order to eliminate the threats in systems, a credibility evaluation model based on fuzzy clustering is presented in this paper, which uses information entropy to determine the optimal number of clusters and utilizes the probability density function as the degree of similarity. And the new evaluation system can be a very good recommendation to eliminate malicious network nodes and deception by comparison with EigenRep algorithm in the simulation.

1 引言

随着网络技术的迅速发展, P2P 网络因其开放性以及节点(peer)的匿名性和自治性等本质特征得到广泛应用, 这些特点为网络技术和网络通信的发展提供了新的发展方向。但是也正是这些特点为计算机病毒、垃圾数据、伪造文件等在 P2P 网络上的传播提供了便利的条件, 如 VBS.Gnutella worm 蠕虫病毒的流行等等。最近 Kazaa 对网络中文件的研究表明, 有超过 50% 的音频文件是被污染的(polluted) [1], 另外, 由于缺乏激励机制, 有 25% 的节点是 free riders [2] (只从其他节点下载文件, 而不提供文件上载服务)。文献[3]使用博弈论模型, 从理论角度分析表明: 引入依据节点的“可信程度”的高低来决定为节点提供何种服务的机制是解决上述两个问题的一个有效的办法。文献[4-6]均给出了各自的信任模型, 并通过仿真实验说明 P2P 网络中引入信任机制可以抑制上述问题。

分布式、匿名性和自治性是 P2P 网络的本质属性,也是其现在在网络上被广泛使用的主要原因。因此,在 P2P 环境下的信任模型要解决的核心问题是:在不损失这些本质属性的前提下,实现节点信任数据的计算、存储和分发,并且上述各环节的实施应占用很少的网络资源,同时保证规模的可扩展性。P2P 环境下的信任模型所面临的首要挑战是协同作弊问题,在恶意节点形成作弊的团体、协同伪造信任值时,信任模型能否有效地识别,乃至遏制作弊行为是评价模型的重要指标。

本文旨在构造一种在 P2P 环境下的一种新的评价体系,建立这种新的评价体系主要是考虑到在传统的信任管理模型中,在计算推荐信任的时候,推荐节点的集合都是和其有过交易历史的节点的集合,但是由于每个节点的兴趣向量和声誉的不同,往往使形成的推荐信任链是不合理的。因为推荐信任链的形成是有条件的,推荐者的选择也是有条件的。本文利用模糊聚类的方法,利用信息熵来确定最优的聚类数目,利用概率密度函数来选择初始聚类中心,引入余弦相似函数作为相似程度。充分考虑网络中节点行为的一致性,很好地解决了在传统的信任管理模型中选择其交易节点和推荐节点的盲目性的缺陷。并且由于聚类的目标评价函数是以网路节点之间的相似度作为评价函数,所以本文期望将诚信节点聚成一类,恶意推荐和协调作弊的节点不发生聚类现象或者只聚成若干个小类。因此节点在选择其交易节点和推荐节点的时候只能选择诚信节点进行交易,网络中节点的信任值都是由诚信节点计算得到,这样就能很好的遏制恶意推荐和协同作弊的行为。

本文结构如下:第二节介绍模糊聚类的定义;第三节讨论聚类数目的确定问题;第四节研究聚类中心的选择;第五节构造一个合适的目标评价函数;第六节阐述在聚类的情況下建立一个新的评价体系的方法,并通过仿真,将新系统和已有 EigenRep 算法进行比较,结果表明新的评价能够很好地消除节点的恶意推荐和欺骗。

2 模糊聚类方法

设有限集 $X = \{x_1, x_2, \dots, x_n\}$, 其中 $x_j \in R^d$ 。模糊聚类算法 FCM (Fuzzy C-Means) 采用误差平方和函数作为聚类准则函数^[8]:

$$J = \sum_{i=1}^m \sum_{j=1}^n \mu_{ij}^h \|x_j - v_i\|^2 \quad (1)$$

式中 n 为样本数, m 为给定的类别数, 并且 $1 < m < n$, h 为加权幂指数, v_i 为第 i 类的聚类中心, μ_{ij} 为样本 j 属于第 i 类的程度, 满足 $\sum_{i=1}^m \mu_{ij} = 1, j = 1, 2, \dots, n$ 。

FCM 算法就是要求使 J 达到最小值的聚类结果。因此将 J 分别对 μ_{ij}, v_i 求导, 令它们的导数为 0, 并代入条件 $\sum_{i=1}^m \mu_{ij} = 1$, 解得:

$$\mu_{ij} = \frac{\left(\frac{1}{\|x_j - v_i\|^2} \right)^{\frac{1}{h-1}}}{\sum_{k=1}^m \left(\frac{1}{\|x_j - v_k\|^2} \right)^{\frac{1}{h-1}}} \quad i = 1, 2, \dots, m; j = 1, 2, \dots, n \quad (2)$$

$$v_i = \frac{\sum_{j=1}^n (\mu_{ij})^h x_j}{\sum_{j=1}^n (\mu_{ij})^h} \quad i = 1, 2, \dots, m \tag{3}$$

FCM 算法是通过对式 (2) 和式 (3) 进行迭代来完成的, 当 J 收敛到极小值时, 就得到了最终的聚类中心, 从而得到最终的聚类结果。

3 聚类数目的确定

熵是用来描述原子分布无序程度的物理量, Shannon 将熵的概念引入到在信息论中, 将信息熵定义为一个信源发出某一消息所含信息量的度量。当某一信源发出的消息越确定时, 该信源的信息熵越小。由于数据点的分布类似于原子的分布, 所以根据信息熵的理论, 当聚类的划分越合理, 数据点在某一聚类上的归属越确定时, 该聚类信息熵值越小。

本文基于上面的信息熵理论, 以平均信息熵的大小作为评判聚类数目的标准。首先确定期望产生的聚类数目范围 $[m_{\min}, m_{\max}]$ 。平均信息熵值的定义为^[11,12]

$$H(k) = - \sum_{i=1}^k \sum_{j=1}^n \left[\frac{\mu_{ij} \times \log_2(\mu_{ij}) + (1 - \mu_{ij}) \times \log_2(1 - \mu_{ij})}{n} \right] \tag{4}$$

式中 μ_{ij} 为样本 j 属于聚类 i 的程度, $\mu_{ij} \in [0,1], \forall i, j = 0, 1, \dots, L$ 。当 k 从 m_{\min} 增加到 m_{\max} 时, 就产生 $m_{\max} - m_{\min} + 1$ 个 $H(k)$, 选取最小的一个 $H(k)$ 所对应的聚类数目 k 作为最终的聚类数目 m 。

利用信息熵来确定最优聚类数目的基本步骤如下:

- ① 设置最大聚类数目 m_{\max} 和最小聚类数目 m_{\min} , 阈值 ε , 并设 $k = m_{\min} - 1$;
- ② 随机初始化隶属矩阵 $U^{(t)}$, $t = 0, \quad k = k + 1$;
- ③ 更新隶属矩阵 $U^{(t)}$ 和聚类中心 $V^{(t)}$, $t = t + 1$;
- ④ 当 $|J^{(t)} - J^{(t-1)}| > \varepsilon$ 时, 返回③;
- ⑤ 计算 $H(k)$, 记下此时聚类数目 k 。如果 $H(m) = 0$, 则 $H(m) = H(k)$; 如果 $H(m) < H(k)$, 则 $H_m(x) = H_k(x)$, 并用当前 k 值更新 m 值; 如果 $k > m_{\max}$, 则 m 即为最终聚类数目; 否则返回②。

4 初始聚类中心的选择

在 FCM 算法的初始化中, 不仅包括初始化聚类数目, 还有初始聚类中心的确定。上面已经用信息熵来确定最优的聚类数目, 下面利用一种概率密度函数来对初始聚类中心进行选择。定义样本点 x_i 处的密度函数如下^[10]:

$$D_i^{(0)} = \sum_{j=1}^N \frac{1}{1 + f_d \|x_i - x_j\|^2} \tag{5}$$

式中 $f_d = 4/r_d^2$, r_d 为邻域密度有效半径, 它的选择与数据集合的分布特性有关。取 r_d 为 N 个

样本的均方根距离的 $\frac{1}{2}$ ，即

$$r_d = \frac{1}{2} \sqrt{\frac{1}{N(N-1)} \sum_{j=1}^N \sum_{i=1}^N \|x_i - x_j\|^2} \quad (6)$$

由式(6)可知，在 x_i 周围样本点越密集，则 $D_i^{(0)}$ 值就越大，所以用它表示在样本空间中样本点的密集程度。

令 $D_1^* = \max\{D_i^{(0)}, i=1, 2, \dots, N\}$ ，对应的 x_1^* 取为第一个初始聚类中心，则第 k 次迭代时的聚类中心的密度函数调整关系式为：

$$D_i^{(k)} = D_i^{(k-1)} - D_k^* \frac{1}{1 + f_d \|x_i - x_k^*\|^2} \quad k=1, 2, \dots, m-1 \quad (7)$$

式中 m 为聚类数目。令 $D_k^* = \max\{D_i^{(k-1)}, i=1, 2, \dots, N\}$ ，对应的样本点 x_k^* 取为第 k 个初始聚类中心位置。式(5)~(7)决定中心初始化方法。

5 目标评价函数的构造

在模糊聚类算法中，模糊聚类效果的评价是通过目标评价函数来衡量的。一种好的目标评价函数既要考虑模糊划分同一类中的紧凑程度又要考虑类与类之间的离散程度。在本文中，为了构造一个合适的目标评价函数，引入相似度的概念。相似度 C_{ij} 刻画了节点 i 和节点 j 的评分行为之间的相似程度，节点 i 和节点 j 之间的相似度越大，说明节点 i 和节点 j 对网络中其他节点的看法越趋于一致，这样节点 i 在选择它的推荐节点的时候就要优先考虑这些和它相似度比较相近的节点^[7]。

在传统的信用评价体系中，节点 i 的推荐节点的集合都是以往和它有过交易历史的节点的集合。在本文中，我们以节点之间的相似度作为评价函数，使用模糊聚类的方法将网络的节点分为若干类，节点 i 的推荐节点的集合就是其所在类里面和其有过交易的节点的集合。

为了构造合适的目标评价函数，我们引入余弦相似函数来评价同一类里面的相似程度：

$$C_{ij} = \frac{\sum_k x_{ik} x_{jk}}{\sqrt{\sum_k x_{ik}^2} \cdot \sqrt{\sum_k x_{jk}^2}} \quad (8)$$

其中， $X_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ 为节点的兴趣向量。 C_{ij} 刻画类中节点之间的相似程度，其值越大，说明聚类的越合理，节点之间的相似度越相近。为了再刻画类与类之间的离散程度，本文引入下面的函数作为评价函数：

$$D_{ij} = \frac{1}{\frac{1}{\sqrt{2\pi}} \exp(-\frac{(r_i - v_j)^2}{2})} \quad (9)$$

其中， r_i 为节点 i 的声誉值， v_j 为聚类中心。 D_{ij} 刻画类之间的离散程度，值越大，说明聚类的结果越合理，同样类与类之间越离散。

基于上面的讨论，本文构造了下面的目标评价函数：

$$f(C_{ij}, D_{ij}) = \sum_{k=1}^m C(k) + D(k) \tag{10}$$

由于 $C(k)$ 和 $D(k)$ 的数量级不一样，所以在 $D(k)$ 前面加一个系数 β 起平衡作用， m 表示最优的聚类数目。

6 基于模糊聚类的信誉评价体系

根据上面的模型，我们选取一个存在的网络，共有 1000 个节点，平均每个节点共有 2000 次的交易历史记录，诚信节点的信誉度 r 服从 $(0.9, 1.0)$ 的均匀分布，诚信节点每次交易完成以后以概率 r 给对方做出客观评价，以 $(1-r)$ 的概率做出随机评价；作弊节点的真实信誉度 r 服从 $(0.0, 0.1)$ 的均匀分布，每次交易完成以后以概率 r 给对方做出客观评价，以 $(1-r)$ 的概率做出随机评价。

6.1 单个节点作弊

从图 1 可以看出，当作弊节点数量不超过 70% 的时候，本文提出的基于模糊聚类的信任评价算法有较好的准确度且一直优于 EigenRep 的信誉评价算法。这是因为在单个节点作弊模型下，所有诚信节点会聚类为一个大的集合，作弊节点因为彼此之间也互相随机进行评价，所以基本上不会发生聚类的现象，每个节点在选择其推荐节点和交易节点的时候基本上选择的都是诚信节点，所以每个节点的信誉评价值主要以诚信节点的观点为主，因此整个网络中信任值都是由诚信节点来决定的，这样就很好的遏制了恶意推荐和协同作弊的行为。仿真结果表明，当网络中存在多个互不联系的小的作弊集合的时候，其仿真结果和单个节点的作弊模型类似。

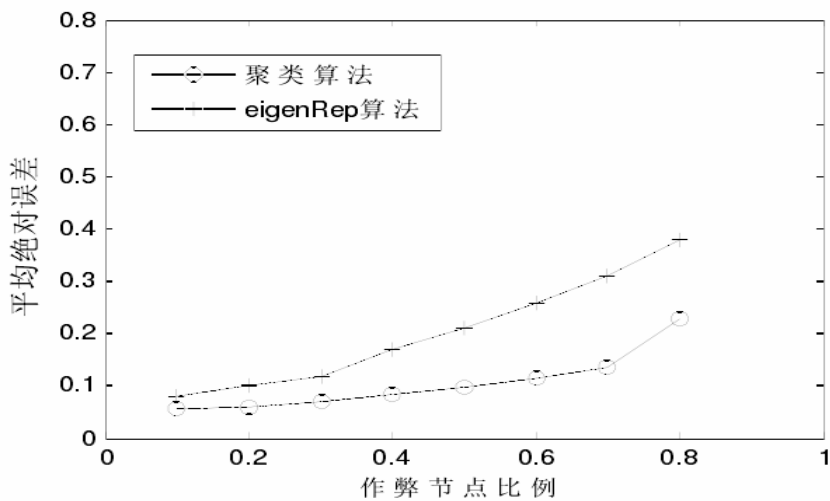


图 1 单个节点作弊

6.2 单组作弊节点

从图 2 可以看出，当作弊节点的比率低于 50% 的时候，基于模糊聚类的信任评价算法要优于 EigenRep 信誉评价算法，当作弊节点比率高于 50% 的时候，模糊聚类算法的误差开始显著增大。这是因为在单组作弊模型下，所有作弊节点互相串通，其行为高度一致，整个网络最终聚类为两个节点集合：诚信节点集合和协同作弊节点集合，并且这两个集合的观点完全相反，此时网络中的节点将无法区分每组节点究竟是诚实节点还是作弊节点，因此节点在选择交易节点和推荐节点的时候将无法做出正确的判断，这样就导致了整个网络中的节点的信任度处于不真实的状况，整个网络处于一个瘫痪的状态。在这种情况下通常较大的节点集合会有较高的信誉度，所以当作弊节点比率超过 50% 的时候，网络中节点的推荐节点的集合可能是作弊节点的集合，使得网络中节点的信誉度的计算不完全由城市节点完全决定，这样就导致了较大的误差，使得误差快速增加。而 EigenRep 因为有桩节点的存在，所以当作弊节点超过 50% 的时候，误差仍然基本呈线性上升的趋势。但是实际的网络环境中很少会遇到所有作弊节点同时串通为一个集团的情况。而在实际情况中，网络中的作弊节点的数目一般是不超过 50%，因此基于模糊聚类的信任评价算法还是要优于传统的信任管理模型。

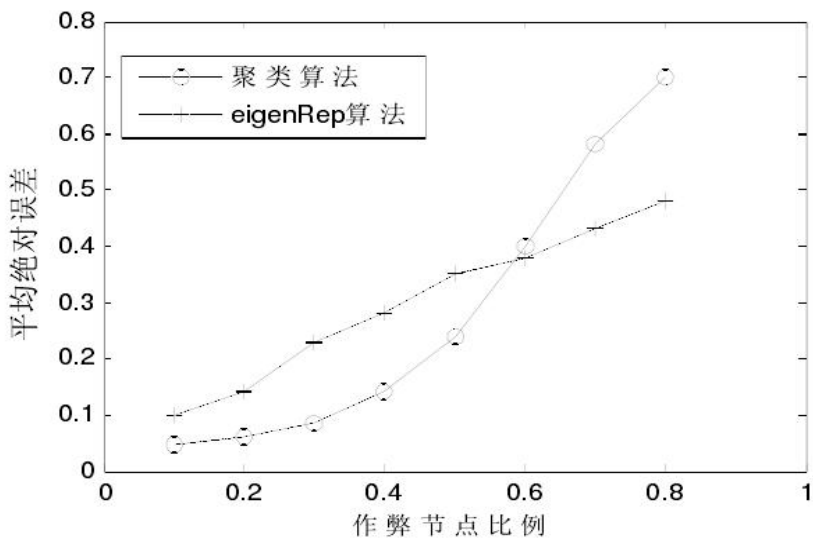


图 2 单组作弊节点

7 结论

根据上面的仿真结果，可以看出每个诚信节点都对对方做出客观评价，其行为具有高度的一致性，因此诚信节点通常都可以聚类成一个较大的节点集合；作弊节点由于作弊方式不同，通常会聚类成若干个小的节点集合甚至不发生聚类现象。经过聚类后可以得到若干个节点集合，因为集合内部节点互相评价较高，所以较大（节点数量较多）的集合在算法初始迭代的时候可能会获得较高的初始信誉度，又因为每个节点集合的信誉度大致按照加权平均的方式计算的，因此初始信誉度较高的节点在以后的迭代计算中一般会继续保持较高的信誉度，

这个最大的集合对其他节点集合的评价也将大幅度影响其他节点的信誉度,可以说整个网络中所有节点的最终信誉度主要是由最大的节点集合决定的。因此利用模糊聚类的方法对初始的网络进行处理,由于每个诚信节点都对对方做出客观评价,其行为具有高度的一致性。所以利用相似度作为评价函数将诚信节点聚为一类,这样网络中每个节点在选择其交易节点和推荐节点都是在诚信节点的类中进行选择,然后再计算节点的信任度,这样就能够很好的减少和消除恶意推荐情况的存在,能够很好地维持网络中节点的信任度的更新和网络的稳定。

参考文献

- [1] Liang J, Kumar R, Xi Y, Ross K. Pollution in P2P file sharing systems. In: Makki K, Knightly E, eds. Proc. of the IEEE Infocom 2005, Vol.2. Miami: IEEE Press, 2005. 1174~1185.
- [2] Saroiu S, Gummadi PK, Gribble SD. A measurement study of P2P file sharing systems. In: Kienzle MG, Shenoy PJ, eds. Proc. Of the Multimedia Computing and Networking 2002 (MMCN 2002). SPIE Press, 2002.
- [3] Buragohain C, Agrawal D, Suri S. A game theoretic framework for incentives in P2P systems. In: Shahmehri N, Graham RL, Carroni G, eds. Proc. of the 3rd Int'l Conf. on Peer-to-Peer Computing (P2P 2003). Los Alamitos: IEEE Press, 2003. 48~56.
- [4] Shi WM, Yang HF, Wu YS, Sun X. Numerical Analysis. 2nd ed., Beijing: Beijing Institute of Technology Press, 2004. 91~93 (in Chinese).
- [5] Xiong L, Liu L. PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. IEEE Trans. on Knowledge And Data Engineering, 2004,16(7):843~857.
- [6] Kamvar SD, Schlosser MT, Garcia-Molina H. The eigentrust algorithm for reputation management in P2P networks. In: Bakonyi P, Hencsey G, *et al.*, eds. Proc. of the 12th Int'l World Wide Web Conf. Budapest: ACM Press, 2003. 640~651.
- [7] 陈锦言. P2P 环境下的评价体系[D]. 天津: 天津大学博士学位论文. 2008, 8.
- [8] Chen Kejia, Ji Ping. Dynamic Advanced Planning And Scheduling With Frozen Interval for New orders[J]. Chinese Journal Of Mechanical Engineering, 2007, 20(4): 117~119.
- [9] 牛强, 夏士雄, 周勇. 等. 改进的模糊 C-均值聚类方法[J]. 电子科技大学学报, 2007, 36(6): 1257~1259.
- [10] Wen Yu, Xiao Ouli. Fuzzy identification using fuzzy neural networks with stable learning algorithms [J]. IEEE TRANSACTIONS ON FUZZY SYSTEMS. 2004, 12(3): 411~420.
- [11] 侯惠芳, 刘素华. 一种改进的基于遗传算法的模糊 C-均值算法[J]. 计算机教程, 2005, 31(17): 152~154.
- [12] 王秀珍. 模糊聚类分析及其应用[J]. 长沙大学学报, 1999, 13(4): 46~49.
- [13] 李景涛, 荆一楠, 肖晓春等. 基于相似度加权推荐的 P2P 环境下的信任管理模型[J]. 软件学报, 2007, 18(1): 157~167.

平面紧凑型双通带滤波器设计

刘艳萍

摘要: 提出了一种新型的双通带滤波器设计方法。该滤波器采用两组谐振器来实现, 其中一组谐振器镶嵌在另外一组谐振器内部, 从而滤波器的尺寸很小。外部的谐振器有两个功能, 一是用于产生第一个通带, 二是用于给内部的谐振器馈入信号。内部谐振器用于产生第二个通带, 交叉指式耦合结构用于实现第二通带级间电磁耦合。对第二通带而言, 该结构可实现源-负载的耦合, 从而在第二个通带附近产生一对传输零点, 提高了滚降特性。整个滤波器采用平面结构, 易于加工和与其他电路集成。滤波器的每一个通带每侧都有至少一个传输零点, 从而获得很高的边缘选择性。

关键词: 双通带滤波器; 内嵌式; 微波; 交叉指电容

Abstract: A planar dual-band bandpass filter based on a novel feed scheme is presented in this paper. The proposed filter employs two sets of resonators operating at diverse frequencies to generate two passbands. A novel scheme is introduced to feed the resonators. One set of resonators is utilized to not only generate lower passband but also feed other resonators. Source-load coupling for upper passband is inherently realized. This scheme provides sufficient degrees of freedom to control the center frequencies and bandwidth of the two passbands. Four transmission zeros can be created close to passband edges, resulting in high skirt selectivity. To validate the proposed idea, a demonstration filter is implemented. The design methodology as well as the experimental results is presented.

1 背景

近年来, 多频带、多标准无线通信系统引起了越来越多的关注, 多频带带通滤波器具有极其广泛的应用。在无线通信系统中, 滤波器为极其重要的组件, 其作用是使必要的信号通过, 并且将不必要的信号滤除。实现双频段信号的筛选, 传统的方法是使用两个单通带滤波器分别对信号进行处理, 每一个滤波器筛选出对应的频段信号, 这种方法的特点是简单可行, 容易实现, 其缺点在于集成度不够, 对于无线通信系统要求小体积、轻重量的目标来说, 实际上并不符合需求。因此如何设计一种能够有效集成的双通带滤波器, 并且通带之间有较好的隔离效果是当前的重要课题。

双通带滤波器的实现方式有很多种, 大概可以分为两类。(1) 由两组谐振器并联实现双通带滤波器。这种实现方式由两组谐振器组成, 采用公共的输入输出端口, 每组谐振器产生一个通带。这种滤波器的缺点是空间体积大, 不易集成。(2) 由一组谐振器来实现双通带滤

波器。这种滤波器中的谐振器同时工作在二个频率上，产生两个通带。其特点是集成度很高，工艺简单，缺点在于两个通带是相互关联的，频率和带宽都难以控制。

2 滤波器设计

该滤波器的结构如图 1 所示，这是一个微带滤波器，包括上层微带结构、中介质基板和底层金属地板。上层微带结构包括输入端口和输出端口，2 个外部谐振器和 2 个内嵌的谐振器。外部谐振器的电长度较长，用于产生低频率通带，即第一通带。内部的谐振器的电长度较短，用于产生高频率通带。内部谐振器通过交叉指电容来进行信号的耦合，交叉指电容的结构如图 2 所示。50 Ω 输入输出端口是直接连接在外部谐振器上。

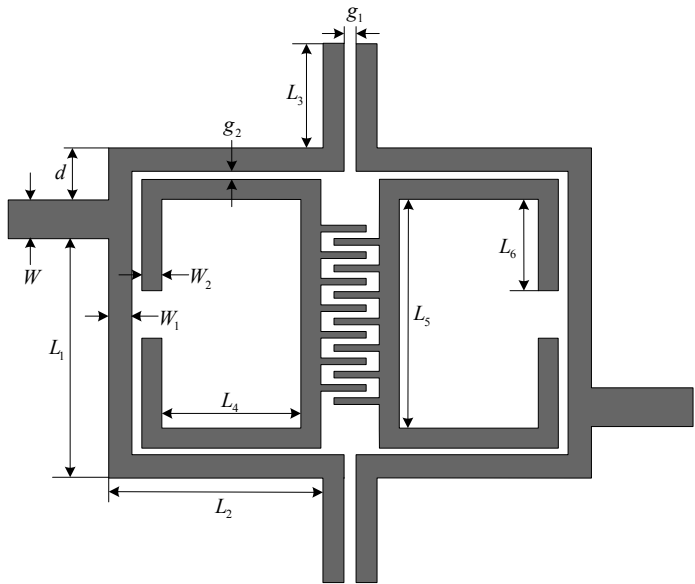


图 1 滤波器的结构

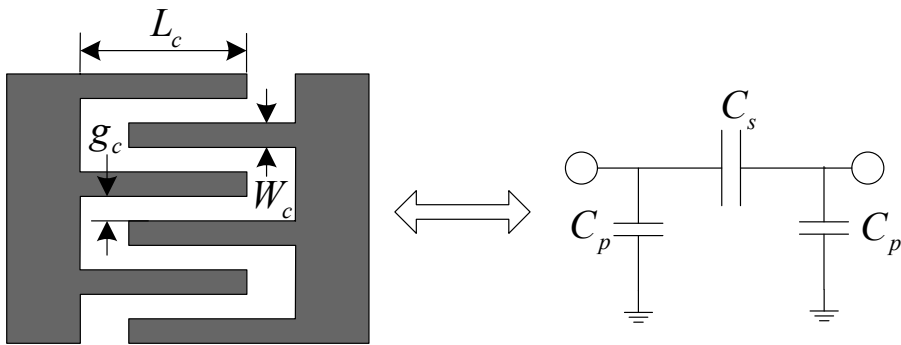


图 2 交叉指结构

该滤波器的馈电和耦合机理如图 2 所示。对于第一个通带而言，输入输出端口直接连接在外部谐振器上，因此信号直接馈入。在第一个通带的中心频率上，只有谐振器 1、4 谐振。谐振器 2、3 不谐振，但是他们对谐振器 1、4 有加载作用，使之频率降低，从而减小的尺寸。在第二个通带的中心频率上，谐振器 1、4 不谐振，在此他们相当于馈电的一部分，用于给内部的谐振器 2、3 进行馈电。由于 1、4 之间存在耦合，因此形成了两个耦合路径，这两个耦合路径可以用于产生一对传输零点，从而提高通带边缘的滚降特性。

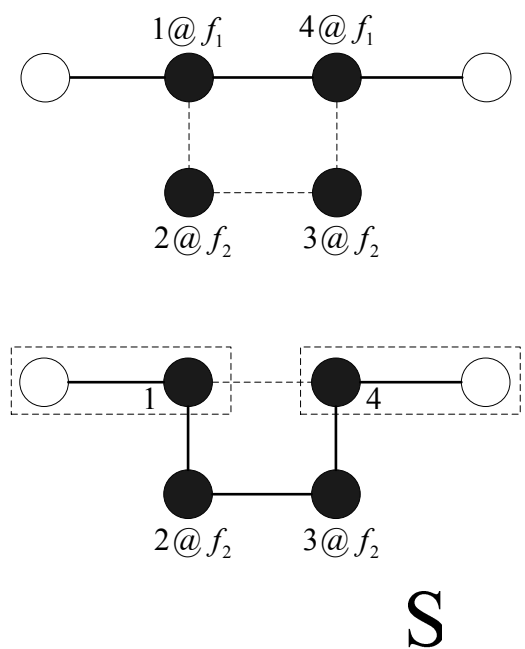


图 3 馈电与耦合机理. (a) 低频通带. (b) 高频通带

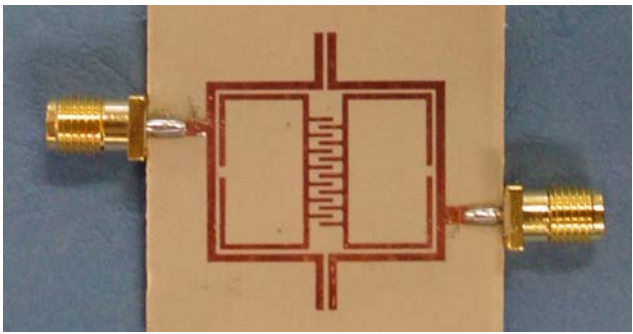


图 4 滤波器照片

根据上述机理，我们设计了一个工作在 DCS 频段和无线局域网频段的双通带滤波器，其照片如图 4 所示。该滤波器的介质基板的介电常数为 2.94，厚度为 0.762mm。微带结构的尺寸为： $L_1=12.5\text{ mm}$, $L_2=10.7\text{ mm}$, $L_3=4.8\text{ mm}$, $L_4=7.6\text{ mm}$, $L_5=14.2\text{ mm}$, $L_6=6.6\text{ mm}$, $W_1=1\text{ mm}$, $W_2=0.7\text{ mm}$, $g_1=0.3\text{ mm}$, $g_2=0.25\text{ mm}$, $W=1.9\text{ mm}$, $L_c=0.9\text{ mm}$, $W_c=0.5\text{ mm}$, $g_c=0.3\text{ mm}$ 。整个滤波器的大小为 $0.22\lambda_g \times 0.27\lambda_g$ ， λ_g 是第一通带谐振频率对应的波长。

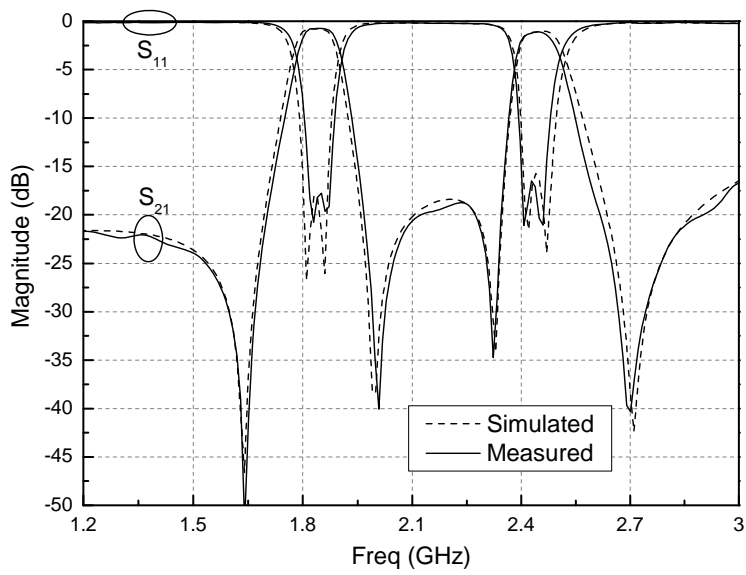


图 5 仿真和测试结果

该滤波器的测试与仿真结果如图 5 所示。由图可见，仿真结果与测试结果非常吻合。第一通带的中心频率是 1.85GHz，为 DCS1800 频段的中心频率，1-dB 带宽为 86 MHz 或者相当于 4.9%，可以满足 DCS 系统的频率要求（1.805 到 1.880 GHz）。插入损耗为 0.8dB，回波损耗大于 15dB，该通带在 1.64GHz 和 2.01GHz 处产生了两个传输零点，极大地提高了边缘选择性。第二通带的中心频率是 2.45GHz，为 2.4GHzWLAN 系统的中心频率，1-dB 带宽为 95MHz，可以满足无线局域网的要求。最小插入损耗为 1.6dB，通带内回波损耗大于 15dB，由于源-负载的耦合作用该通带在 2.32GHz 和 2.70GHz 处产生了两个传输零点，它们处于通带的边缘位置，提高了滤波器的滚降系数。

从频率响应可以看出，该滤波器可以工作在 DCS 和 WLAN 频段上，可用于筛选这两个个系统的有用信号并抑制干扰信号，该滤波器有 4 个传输零点，滚降特性好，插损较小。

3 总结

本文提出了一种新型的双通带滤波器设计方法并用实验进行了验证。该滤波器共实现了 4 个传输零点，每个通带每侧至少有一个传输零点，提高了滤波器的滚降特性。该滤波器电路为平面结构，谐振器之间采用内嵌式的结构，整个电路尺寸小，结构紧凑，空间利用率高，容易加工。本滤波器的地面是完整的地，可以有效地防止信号泄露，并且易于和其他微带电路集成。

参考文献

[1] J.-T. Kuo, T.-H. Yeh, and C.-C. Yeh, "Design of microstrip bandpass filter with a dual-passband response," IEEE Trans. Microw. Theory Tech., vol.53, no.4, pp.1331-1337, Apr. 2005.

[2] S. Sun and L. Zhu, "Compact dual-band microstrip bandpass filter without external feeds," IEEE Microw.

Wireless Commun. Lett., vol.15, no.10, pp.644-646, Oct. 2005.

- [3] X. Y. Zhang and Q. Xue, "Novel centrally loaded resonators and their applications to bandpass filters," IEEE Trans. Microw. Theory Tech., vol. 56, no. 4, pp. 913-921, Apr. 2008.
- [4] R. Cameron, "Advanced coupling matrix synthesis techniques for microwave filters," IEEE Trans. Microw. Theory Tech., vol.51, no.1, pp.1-10, Jan. 2003.
- [5] J. S. Hong, and M. J. Lancaster, Microwave Filter for RF/Microwave application. New York: Wiley, 2001.

作者简介

刘艳萍，女，1975年出生，2005年6月华南理工大学电子与信息学院硕士研究生毕业，现为中国人民解放军75708部队工程师，主要从事智能天线，射频与无线通信等方面的研究。

一种基于用户背景知识的文本聚类方法

沈志辉 袁再江

(1. 第二炮兵工程学院, 陕西西安, 710025; 2. 第二炮兵装备研究院, 北京, 100085)

摘要: 传统的文本聚类模型没有考虑到用户背景知识和经验, 本文提出了一种基于用户背景知识的文本聚类方法。首先运用向量空间模型表示文本特征信息, 然后引入一阶谓词逻辑来表达用户背景知识, 并通过关联规则挖掘出主题频繁项集, 相关联的文本集即为聚类中所包含的初始文本, 最后通过二进制相似性度量进行文本集的合并, 最终实现文本聚类。实验结果表明该方法是有有效的。
关键词: 用户背景知识; 文本聚类; 关联规则

A Method for Documents Clustering Based on User Background Knowledge

SHEN Zhi-hui YUAN Zai-jiang

(1the Second Artillery Engineering Institute, Xi'an 710025, China;
2 High Technology Institute of the Second Artillery, Beijing 100085, China)

Abstract: The user background knowledge and experience hasn't been considered in traditional documents clustering models. A method for documents clustering based on user background knowledge is proposed in this paper. First, the vector space model is used to represent the text feature information ,and then by using first order predication logic, the technology of background knowledge representation is presented. A frequent item sets can be found by using the association rules discovery algorithm, corresponding documents can be seen as initial clusters. These clusters are merged according to binary similarity between clusters. Experimental results show the effectiveness of this document clustering method.
Keywords: User Background Knowledge; Documents Clustering; Association Rules

1 引言

信息资源的日益丰富和繁杂增加了人们获取所需信息的难度, 文本聚类分析作为一种有效的文本数据挖掘手段, 已经成为信息处理领域中的一项重要研究课题。
文本聚类是指把文本集合按照相似性自动归成若干个称为簇或者类的子集, 使得每个簇中的文本之间具有较大的相似性, 而簇之间的文本具有较小的相似性。目前, 关于文本聚类

典型的方法有:层次聚类法 (agglomerative hierarchical clustering, AHC) [1,2], 如在 Stanford 大学数字图书馆系统中的 Soina 系统; 平面划分法[3,4], 以 K-Means 算法为代表; 基于 SOM (Self-Organizing Maps) 神经网络法[5, 6,7]等, 但这些方法都没有结合用户的兴趣、经验等背景知识, 如何把用户的背景知识融入文本聚类分析过程是当前研究的一个难题和热点。

本文综合考虑了用户背景知识, 建立一种基于用户背景知识的文本聚类模型。

2 基于用户背景知识的文本聚类模型

2.1 文本特征表示

文本表示最常见的方法是基于向量空间模型 (Vector Space Model) 的方法, 其基本思想是: 把文本表征成由特征项构成的向量空间中的一个点, 通过计算向量之间的距离, 来判定文本之间的相似程度。

应用向量空间模型建立主题特征向量, 然后建立文本的主题向量。定义主题特征向量: $T_i = [(k_{i,1}, w_{i,1}), (k_{i,2}, w_{i,2}), \dots, (k_{i,l}, w_{i,l})]$, 其中, $k_{i,j}$ 代表主题 T_i 中的第 j 个关键字/短语, $w_{i,j}$ 为第 j 个关键字/短语 $k_{i,j}$ 对应的权值, 表示该关键字/短语在该主题中的重要程度, $\sum w_{i,j} = 1$, $1 \leq j \leq l$, $l = \|T_i\|$, 为主题 T_i 中关键字/短语的个数。

设 D 是文本的集合, 其中每一个文本 $D_j \in D$, 文本主题向量定义为 $D_j(T_i) = [\mu_{i,1}^j, \mu_{i,2}^j, \dots, \mu_{i,l}^j]$, $\mu_{i,k}^j = \frac{\|UK_{i,k}\|}{\|UD_j\|} \cdot w_{i,k}$, 其中 $\|UK_{i,k}\|$ 是主题 T_i 的第 k ($1 \leq k \leq l$) 个关键字/短语 $k_{i,j}$ 在 D_j 中出现的频度, $\|UD_j\|$ 为 D_j 中有效词的个数, $w_{i,k}$ 为第 k 个关键字/短语 $k_{i,j}$ 在主题特征向量 T_i 中的权值。

计算文本 D_j 和主题 T_i 之间的关联度 $\lambda_{i,j}$: $\lambda_{i,j} = \sum_{k=1}^l \mu_{i,k}^j$, 建立文本—主题事务矩阵:

$$w_{n,m} = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & \dots & \lambda_{1,j} & \dots & \lambda_{1,m} \\ \lambda_{2,1} & \lambda_{2,2} & \dots & \lambda_{2,j} & \dots & \lambda_{2,m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_{i,1} & \lambda_{i,2} & \dots & \lambda_{i,j} & \dots & \lambda_{i,m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \lambda_{n,1} & \lambda_{n,2} & \dots & \lambda_{n,j} & \dots & \lambda_{n,m} \end{bmatrix}$$

并对行向量进行归一化处理得到 $w'_{n,m}$, 使关联度之间具有可比性。

2.2 用户背景知识表示

不同的用户有不同的兴趣, 关心的话题也不一样。同时, 对于某一具体文本, 从不同的角度分析可以聚类为不同的类别。从用户角度来考虑, 把用户倾向、兴趣视为一种用户背景知识, 为了结合背景知识进行文本聚类, 首先要完成用户背景知识表示。

一阶谓词逻辑作为一种形式语言系统, 适合于表示事物的状态、属性、概念等事实性的

知识,也可以用来表示事物之间确定的因果关系。用户背景知识实际上是描述其关心的对象,并形成约束关系,所以采用一阶谓词逻辑能较好的描述文本聚类的用户背景知识。根据用户的兴趣、认识等背景知识,定义相关谓词,并指出每个谓词的确切含义,然后连接有关的谓词,形成一个谓词公式,表达了一条完整的用户背景知识。

根据文献^[8],采用 $Interesting(f(r))$ 、 $Support(f(r),k)$ 、 $Interested(f(r))$ 等谓词来描述用户背景知识。定义如下谓词表达式:

(1) $Interesting(f(r))$, 表示用户直接给出感兴趣的项目集。

(2) $Support(f(r),k) \rightarrow Interesting(f(r))$, 表示在以往的历史挖掘中,如果项目集 $f(r)$ 的支持度大于 k , 那么在新的挖掘中它也是用户感兴趣的模式集。

(3) $Interested(f(r)) \rightarrow Interesting(f(r))$ 表示如果 $f(r)$ 在以往的历史挖掘中是用户感兴趣的模式集,那么在新的挖掘中它也是用户感兴趣的模式集。

2.3 基于用户背景知识的文本聚类模型

根据文本—主题事务矩阵,加入用户背景知识约束规则,通过关联规则挖掘得到主题频繁项集,每个频繁集都关联着一组文本,把这若干组文本作为聚类的候选基类,然后依据基类的二进制相似性度量进行合并类,最后得到文本聚类结果。具体计算步骤如下:

(1) 按照 2.2 节所述建立用户背景知识的谓词表达,形成用户感兴趣的模式 Y 。

(2) 根据 2.1 节所述计算得到文本—主题事务矩阵,对行向量进行归一化处理得到 $w'_{n,m}$ 。

(3) 进行关联规则挖掘得到主题频繁项集^[9]。

① 第一次扫描事物矩阵,即 $k=1$,对于任一项目集 M ,与用户感兴趣的模式 Y 结合,若扩展后的模式是非频繁模式,则将其约减掉,得到符合约束条件的模式。

② 进行 $k(k \geq 2)$ 次扫,先任选两个模式,如果前 $k-1$ 项完全相同,第 k 项不同,则将其组合成 $k+1$ 项,再判断它们是否包含有相同的用户背景知识,如果是,则将用户的背景知识同前面生成的 $k+1$ 项组合,生成候选频繁模式;如果选出的两个模式中前 k 项完全相同,但包含着不同的用户背景知识,则将两个用户背景知识合并,再与前 k 项组合,生成候选频繁模式。

③ 重复步骤②,直到没有新的频繁模式生成。

(4) 根据上述主题频繁项集得到对应的文本集,进行合并操作。对于文本集 p 、 q , n_p 、 n_q 分别表示文本集内文本数, $n_{p \cap q}$ 表示两文本集中共同的文本个数,当 $\frac{n_{p \cap q}}{n_p} > 0.5$, 且

$\frac{n_{p \cap q}}{n_q} > 0.5$ 时,认为两文本集的二进制相似性度量为 1,需合并,否则为 0,不合并。

经过上述步骤得到最终文本聚类结果。

3 试验结果

为了验证本文提出的文本聚类模型,在 Google 进行了试验,评价指标^[7]为“类内准确率” p 和“聚全率” q , 分别定义为

$$p = \frac{\text{类C中符合用户意向文档数}}{\text{类C中文档总数}}, q = \frac{\text{类C中符合用户意向文档数}}{\text{搜索引擎返回所有符合用户意向的文档数}}$$
采用本文方法和 K-均值算法进行聚类分析，结果如下表

表 1 两种方法试验结果比较

	试验 1	试验 2	试验 3	试验 4	试验 5	试验 6	试验 7	试验 8
p_1	0.71	0.82	0.75	0.68	0.77	0.54	0.84	0.78
p_2	0.52	0.64	0.63	0.57	0.51	0.66	0.68	0.65
q_1	0.63	0.67	0.62	0.58	0.74	0.61	0.75	0.65
q_2	0.61	0.58	0.65	0.67	0.51	0.49	0.74	0.70

其中， p_1 、 q_1 表示本文方法试验结果， p_2 、 q_2 表示 K-Means 聚类算法试验结果。

从实验结果可以看出，本文提出的方法聚类效果要优于 K-Means 聚类算法，尤其是在准确率方面更为明显，这是因为本文提出的方法充分考虑到了用户的兴趣、倾向等背景知识，增强了聚类结果的针对性。进一步分析，逐次增加需要聚类的文本数量，两种方法的时间性能比较如图 1 所示：

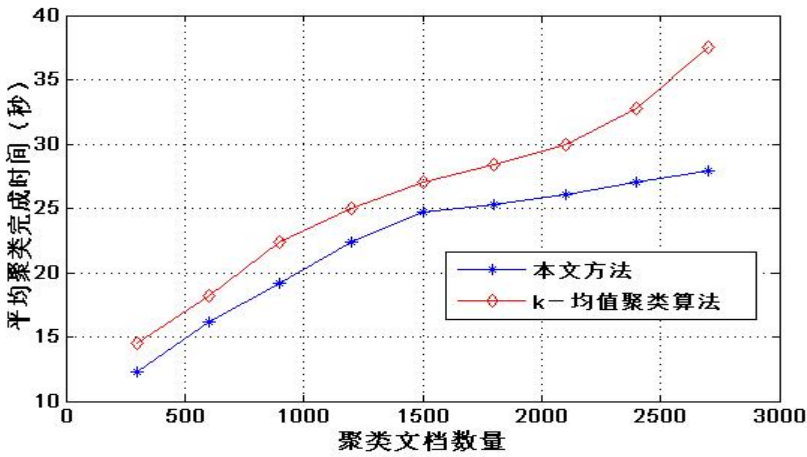


图 1 两种聚类方法时间性能对比

4 结束语

本文提出了一种基于用户背景知识的文本聚类方法，采用谓词逻辑来表示背景知识，通过关联规则挖掘出主题频繁项集，并依据基类的二进制相似性度量进行合并，最终实现文本聚类。试验结果表明该方法具有较好的效果和性能。

参考文献

[1] Jung, Y., Design and evaluation of clustering criterion for optimal hierarchical agglomerative clustering[D]. Phd. Thesis .University of Minnesota, 2001.

- [2] Willett P. ,Recent Trends in Hierarchic Document Clustering: A Critical Review[J]. Information Processing and management, 1988, 24(5): 577-597
- [3] 唐春生, 金以慧. 基于聚类特性的大规模文本聚类算法研究[J]. 计算机科学, 2002. 29(009): 13-15
- [4] Likas, A., N. Vlassis and J.J. Verbeek , The global k-means algorithm[J]. Pattern Recognition, 2003. 36(1).
- [5] Vesanto J, Alhoniemi E. Clustering of the self-organizing map[J]. IEEE Transactions on neural networks, 2000, 11(3): 586-600
- [6] Kohonen T.The self-organizing map[J]. Neurocomputing, 1998, 21(1-3): 1-6
- [7] 陈福集, 杨善林. 一种基于 SOM 的中文 Web 文本层次聚类方法[J].情报学报, 2002. 21(002): 173-176.
- [8] 赵旭俊, 张继福. 基于背景知识的关联规则挖掘算法研究[J]. 通讯和计算机, 2005. 2(006): 11-18.
- [9] 李营, 王儒敬, 王大为等. 基于用户兴趣的搜索结果动态聚类算法[J].计算机工程与应用, 2008. 44(004): 187-189.

作者简介

沈志辉 (1979 -), 男, 博士研究生, 湖南常德, 主要研究方向知识管理、辅助决策。

对一类非线性系统基于三角反馈的Hopf分岔控制

童 炜¹ 万里红²

(1. 中国计量学院机电工程学院, 杭州, 310018;

2. 中国计量学院计算机应用技术学科, 杭州, 310018)

摘 要: 考虑了一类自治系统在三角函数反馈控制下, 实现对 Hopf 分岔点的控制, 通过调节控制器 $u = k \sin(\alpha x + \beta y)$ 中的三个参数, 能很灵活地控制分岔点, 具有工程意义上的实际应用价值。

关键词: Hopf 分岔; 非线性控制; 极限环

1 引言

在我们身边的众多系统中, 具有非线性性的系统占据了绝大多数, 也因此使得科学工作者持续不断的在非线领域中的研究。而非线性系统的分岔研究是非线性动力学的一个重要的前沿课题。此外, 系统的稳定性又是系统能正常工作的基本要求, 在工程上一般要求系统是渐进稳定的。但是在实际中, 系统数学模型中的一些常数常常随着外界干扰、器件老化、环境的变化等一系列因素的影响而发生变化。这些变化如果发生在非线性系统中, 就会使系统的结构稳定性(也即拓扑结构)随着这些参数的变化发生变化, 从而导致分岔现象的发生。但分岔现象又非是一文不值、完全不可取的, 在某些时候, 我们还需要合理的利用分岔现象, 这里就不一一举例了。正因为不同的系统、不同的实际状况下, 分岔现象即能被人们所利用又会影响系统的正常运行, 所以就需要我们能够有效的控制分岔现象。文献^[1]综述了非线性动力学, 特别是分岔和混沌的主要概念和研究现状, 简要地评价了该领域的发展趋势。文献^[2]回顾了近年来分岔控制的研究工作, 展望了这一领域的研究方向, 着重描述了分岔控制在非线性系统镇定问题中的原理和方法。目前分岔控制的主要研究内容有: 将原系统固有的分岔行为进行有效的延迟^[3]; 设计参数值, 使之产生新的分岔^[4-6]; 设计参数值, 改变平衡点的位置^[7]; 改变原非线性系统的拓扑结构, 改变分岔类型^[8]; 改变原系统极限环的幅值、频率^[9,10]等等。

分岔控制研究从一开始就遇到如何应用常规技术的困难, 因为这些控制方法通常都是简单地把整个分岔现象清除掉, 从而达不到镇定或者改变其动力性态的目的。现在据我们了解, 分岔控制只可以用不多的控制方法, 这些方法都有其理论分析和实验或模拟验证, 并且在工程、物理、军事等领域都有一些非常规的应用。在系统中加入适当的非线性参数控制器, 完全可以消除鞍结分岔等一系列破坏性的动力学行为。本文的主要任务就是设计这样一个控制器, 实现开环控制的策略, 对一个系统产生的分岔点进行控制, 同时也实现对产生 Hopf 分岔的控制, 包括在给定的参数区间内, 消除 Hopf 分岔, 或者通过开环控制的增益来改变极限环的幅值。亦即讨论了三种 Hopf 分岔控制问题: 移动 Hopf 分岔、产生新的 Hopf 分岔以及消除

Hopf 分岔。

2 产生Hopf分岔的条件

分岔控制的思想是通过新的控制器来延迟分岔的发生，或者把不稳定的分岔变为稳定的分岔（即所谓的软化分岔），或者扩大系统的安全运行范围，以此来充分利用和提高系统的工作效能。应用在宇航领域，分岔控制很有可能用来充分利用和发挥飞船喷气机的最大潜力，以实现少燃料远距离的太空航行。

为了较好的发展分岔控制，我们有必要重述一下针对连续情形 Hopf 分岔的经典判据。首先考虑如下一个二维时间连续参数化的自治系统：

$$\begin{cases} \dot{x} = f(x, y, \mu) \\ \dot{y} = g(x, y, \mu) \end{cases} \quad (1)$$

这里 $f, g \in C^1(R^2)$ ， $\mu \in R$ 是实数范围上的参变量，假设系统有平衡点 (x^*, y^*) ，即对于所有的 $\mu \in R$ 都有 $f(x^*, y^*, \mu) = g(x^*, y^*, \mu) = 0$ 。令 $J(\mu)$ 是它在平衡点处的雅可比矩阵，且它的特征值为 $\lambda_{1,2} = \alpha(\mu) \pm i\beta(\mu)$ 。可见，随着 μ 的变化， λ 也随着改变， $\lambda_{1,2}$ 从左半平面移到有半平面。假设当 $\mu = \mu^*$ 时，特征值在虚轴上。此时 $\alpha(\mu) = 0$ ， $\beta(\mu) \neq 0$ ，并且 $\frac{\partial \alpha(\mu)}{\partial \mu} \neq 0$ （该

条件保证了特征值有穿越虚轴，也称为特征值在虚轴相交的横截条件，特征值并不与虚轴相切），也就是说，系统在点 (x^*, y^*, μ^*) 处发生了 Hopf 分岔，这个点也称为分叉点。

因此，问题的关键在于设计一个控制器 $u(x, y, \mu)$ ，假如我们并不希望在加入新的控制后会改变原系统平衡点，所以对所有的 μ ，都得满足 $u(x^*, y^*, \mu) = 0$ 。这里，我们假设将控制器加入到第二个方程^[1]，即：

$$\begin{cases} \dot{x} = f(x, y, \mu) \\ \dot{y} = g(x, y, \mu) + u(x, y, \mu) \end{cases} \quad (2)$$

虽然控制器不改变原系统在 (x^*, y^*) 处的平衡点位置，但是可以把 Hopf 分岔点 (x^*, y^*, μ^*) 移动到一个新的位置 $(x^0, y^0, \mu^0) \neq (x^*, y^*, \mu^*)$ ，这是可以实现的。

在系统的平衡点未被改变的情况下，即 $(x^0, y^0) = (x^*, y^*)$ ，那么加入控制器后的系统在 (x^0, y^0) 处的雅可比矩阵为：

$$J(\mu) = \begin{bmatrix} f_x & f_y \\ g_x + u_x & g_y + u_y \end{bmatrix}_{x=x^0, y=y^0} \quad (3)$$

其特征多项式为：

$$\det(\lambda - J(\mu)) = \lambda^2 - (f_x + g_y + u_y)\lambda + f_x(g_y + u_y) - f_y(g_x + u_x) = 0 \quad (4)$$

所以：

$$\lambda_{1,2} = \frac{1}{2}(f_x + g_y + u_y) \pm \frac{1}{2}\sqrt{(f_x + g_y + u_y)^2 - 4[f_x(g_y + u_y) - f_y(g_x + u_x)]} \quad (5)$$

为了表述清楚，以上数值都是在平衡点 (x^0, y^0) 处的值。

根据本节开头的分析,再联系上述特征值,我们可以得到一般二阶单参数非线性系统在平衡点处发生 Hopf 分岔的条件:

a. (x^0, y^0) 是受控系统的平衡点,所以对 $\forall \mu \in R$, 满足:

$$\begin{cases} \overset{g}{x} = f(x^0, y^0, \mu) = 0 \\ \overset{g}{y} = g(x^0, y^0, \mu) + u(x^0, y^0, \mu) = 0 \end{cases} \quad (6)$$

b. 在 (x^0, y^0, μ^0) 处, 系统产出 Hopf 分岔, 所以此时系统雅可比矩阵的特征值为纯虚根, 即:

$$\begin{aligned} (f_x + g_y + u_y) \Big|_{\mu=\mu^0} &= 0; \\ f_x(g_y + u_y) - f_y(g_x + u_x) \Big|_{\mu=\mu^0} &> 0; \\ (f_x + g_y + u_y)^2 - 4[f_x(g_y + u_y) - f_y(g_x + u_x)] \Big|_{\mu=\mu^0} &< 0. \end{aligned} \quad (7)$$

c. 特征值横截穿越虚轴, 即:

$$\frac{\partial \operatorname{Re}\{\lambda_{1,2}(u)\}}{\partial \mu} \Big|_{\mu=\mu^0} = \frac{\partial (f_x + g_y + u_y)}{\partial \mu} \Big|_{\mu=\mu^0} \neq 0 \quad (8)$$

注意到, 以上的控制方法得到普遍应用, 对于这类动力系统, 根据条件(6)~(8)来设计不同的控制器, 能达到不同的控制效果。本文就将采用以下控制器:

$$u = k \sin(\alpha x + \beta y) \quad (9)$$

首先通过 \sin 这个函数能满足不改变原系统平衡点的条件, 另外, 在 x, y 有一个不可测量的时候, 只需将 α 或者 β 设为 0 即可, 具有普遍意义。

3 通过 $u = k \sin(\alpha x + \beta y)$ 对一类函数的分岔控制研究

本文考虑以下系统:

$$\begin{cases} \overset{g}{x} = y \\ \overset{g}{y} = -x + \mu(1 - x^2)y - y^3 \end{cases} \quad (10)$$

显然, 它有唯一的一个平衡点 $(x^*, y^*) = (0, 0)$, 在该点处的雅可比矩阵的特征多项式为:

$$\lambda^2 - \mu\lambda + 1 = 0 \Rightarrow \lambda_{1,2} = \frac{\mu}{2} \pm \frac{1}{2}\sqrt{\mu^2 - 4} \quad (11)$$

当 $\mu > 0$ 时, 系统渐近趋向于一个稳定的极限环; 当 $\mu = 0$ 时, 系统产生稳定的极限环; 当 $\mu < 0$ 时, 系统以 $(0, 0)$ 点为中心稳定型细焦点。可见 $\mu = 0$ 是系统 (10) 的一个分岔点, 如下图所示:

加入控制 $u = k \sin(\alpha x + \beta y)$ 后, 系统 (10) 写成:

$$\begin{cases} \overset{g}{x} = y \\ \overset{g}{y} = -x + \mu(1 - x^2)y - y^3 + k \sin(\alpha x + \beta y) \end{cases} \quad (12)$$

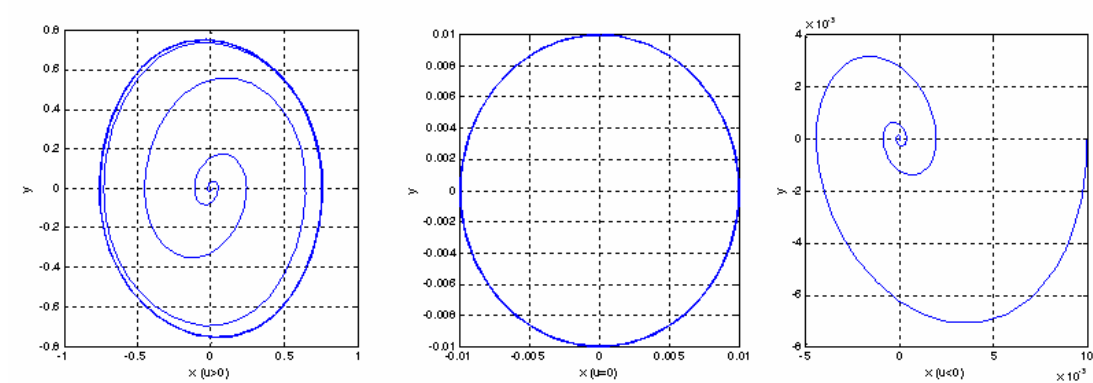


图 3.1 $\mu > 0, \mu = 0, \mu < 0$ 时的解相图

为使系统 (12) 的平衡点仍旧只有 (0,0) 这一点, 故需满足条件 (6):

$$\begin{cases} y = 0 \\ -x + \mu(1 - x^2)y - y^3 + k \sin(\alpha x + \beta y) = 0 \end{cases} \quad (13)$$

方程组(13)只有零解。上述问题等价于方程 $-x + k \sin(\alpha x) = 0$ 只有唯一的零解。故考虑

方程组 $\begin{cases} y_1 = x \\ y_2 = k \sin(\alpha x) \end{cases}$ 的两条轨线只相交于 (0,0) 点。

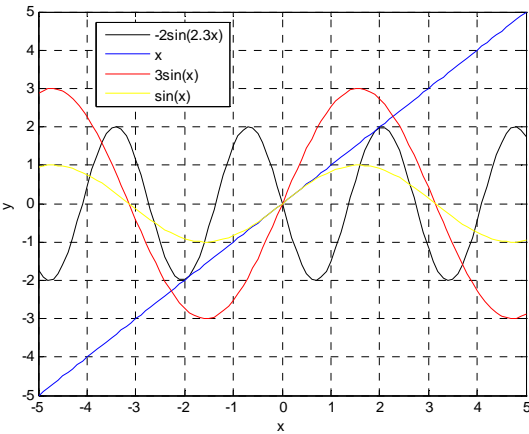


图 3.2 轨线的交点数对照

通过计算, 当 k, α 满足关系式

$$-\frac{3}{2}\pi \leq k\alpha \leq 1 \quad (14)$$

时, 平衡点位置不变。基于这点, 我们重新来算系统(12)在 (0,0) 点处的雅可比矩阵的特征多项式为:

$$\lambda^2 - (\mu + \beta k)\lambda + (\alpha k - 1) = 0 \Rightarrow \lambda_{1,2} = \frac{1}{2}(\mu + \beta k) \pm \sqrt{(\mu + \beta k)^2 - 4(\alpha k - 1)} \quad (15)$$

比较式 (11) 以及对比第 2 节中分析 Hopf 分歧发生的条件, 清楚的可以发现, 当 $k \neq 0$ 即有三角输入时, 通过改变 β 值, 在满足条件 (14) 的前提下, 我们可以任意移动分歧点发生

的位置 $\mu^* = -\beta k$ 。这在工业工程中具有一定实际意义，譬如：在电力系统中，假如在原来的分岔点处会发生电压崩溃事故，我们只要在原来的事故临界点处安置事故危险报警装置，在适当延迟分岔发生的时候，就为技术工人留下一定的时间调节系统，防止事故的发生。

4 仿真检验

这里假设将原来的分岔点 $(x^0, y^0, \mu^0) = (0, 0, 0)$ 移动到 $(x^*, y^*, \mu^*) = (0, 0, 1)$ 处，此时 $\beta k = -\mu = -1$ 。选取 $k = -2, \beta = \frac{1}{2}$ ，再根据条件 (14)，取 $\alpha = 1$ ，即添加控制器后的系统为：

$$\begin{cases} \dot{x} = y \\ \dot{y} = -x + \mu(1 - x^2)y - y^3 - 2\sin(x + 0.5y) \end{cases} \quad (16)$$

系统仿真结果如下：

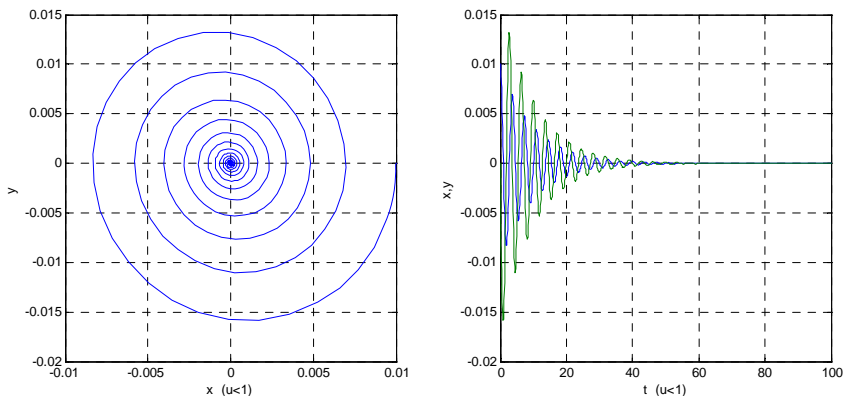


图 4.1 $\mu < 1$ 时的系统相图（左）与时间响应图（右）

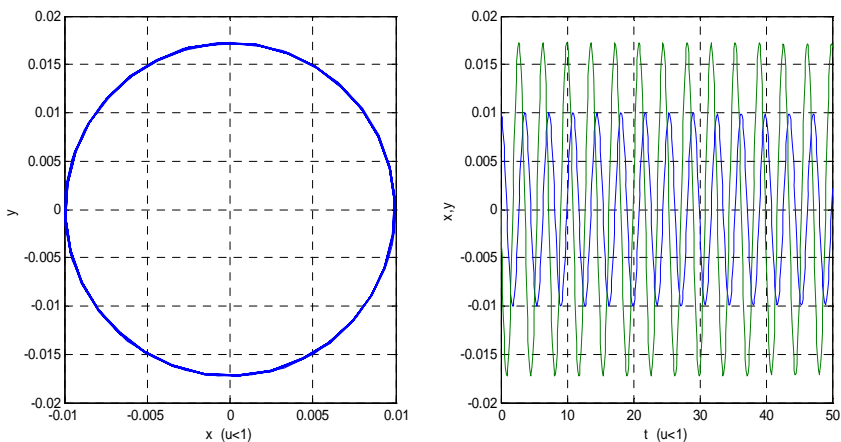


图 4.2 $\mu = 1$ 时的系统相图（左）与时间响应图（右）

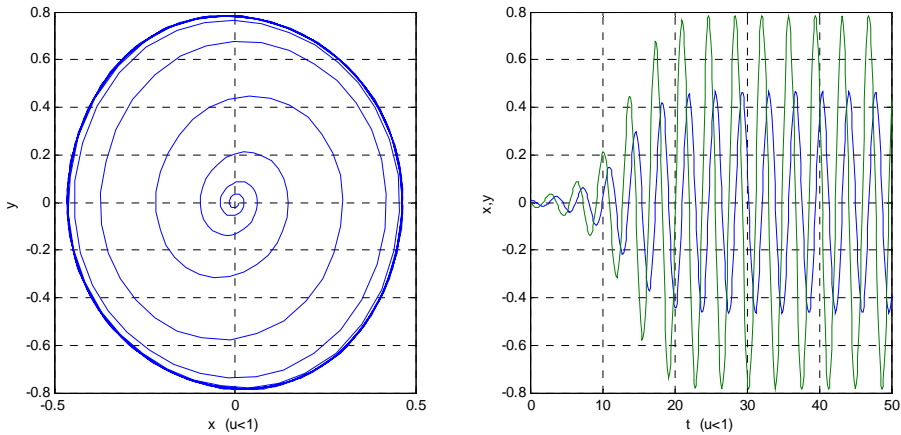


图 4.3 $\mu > 1$ 时的系统相图（左）与时间响应图（右）

5 结论

本文通过三角函数控制器 $u = k \sin(\alpha x + \beta y)$ 有效的改变了分岔点发生的位置，并且，离分岔点越近，收敛速度越慢。这些都可以通过条件控制器的三个参数来实现。相比以往的线性与非线性组合控制器，三角控制器在实现上更容易，可调节的参数更多更灵活。仿真显示，理论分析和实验结果完全一致。

参考文献

[1] 非线性振动、分叉和混沌的最新进展与展望[A].陈予恕.一般动力学(动力学、振动与控制)最新进展[C].北京: 科技出版社, 1994.19~29.

[2] 分叉控制研究综述[A].杨明.信息与控制.2004.4 第 33 卷第 2 期 191~196.

[3] Analysis of static and dynamic bifurcation from a feedback systems perspective [J].Colantonio M C & Donate P D, Dyn.Stab.Syst.1997, 12:293-317.

[4] bifurcations and chaos in a linear control system with saturated input [J].Alvarez J&Curiel L E, Int. J. bifurcation and chaos,1997,7:1811-1822.

[5] Bifurcation and chaos in the Duffing Oscillator with a PID controller[J].Cui F, Chew C H, Xu J&Cai Y, Nonlin.Dynam,1997,12:251-262

[6] Chaotification via arbitrarily small feedback controls; theory, method, and application[J].Wang X F&Chen G, Int.J.of bifur.Chaos,2000,10:549-570.

[7] Bifurcation control: theories, methods, and applications[J].Chen G, Moiola J L, &Wang H O, Int. J. Bifurcation and Chaos,2000,10:511-548.

[8] Bifurcation control of a chaotic system[J].Wang HO&Abed EH, Automatic,1995,31:1213-1226.

[9] Controlling limit cycles and bifurcation, in Controlling Chaos and Bifurcations in Engineering Systems[J]. Calandririni G, Paolini E, Moiola J L&Chen G, ed. Chen G(CRC Press,Boca Raton,FL),1999,200-227.

- [10] Stability of periodic orbits controlled by time-delayed feedback[J],Bleich M E&Socolar J E S, Phys.Lett,1996,A210:87-94.
- [11] 一种统一的状态反馈方法控制 Hopf 分岔[J].方锦清.广西师范大学学报.第 18 卷第 1 期, 2000 年 3 月.

作者简介

童炜 (1984—), 男, 浙江嵊州人, 研究方向: 非线性系统的分岔控制。

万里红 (1986—), 男, 江西新余人, 研究方向: 图像处理与模式识别。

基于属性相似性计算的空间关联规则提取技术研究

王海涛^{1, 2}

(1. 信息工程大学 测绘学院, 河南 郑州 450052;
2. 61363 部队四室, 陕西 西安, 710054)

摘 要: 空间关联规则提取是空间数据挖掘的重要内容, 本文在分析空间实体属性相似性的基础上, 将空间实体属性相似性分为属性名称相似性和属性语义相似性两种, 借鉴序列相似性计算方法, 给出了属性名称相似性计算方法, 并将其应用在空间实体空间关联规则提取当中, 该提取方法在的广泛性、适用性上都有一定的提高。
关键字: 属性相似性; 空间关联规则

The Study of Techology for Association Rules Distilling Based on similarity of Attribute

WANG Hai-tao^{1, 2}

(1. Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450052, China; 2. 61363Troops, Xi'an, 710054, China)

Abstract: The distilling is the important content of spatial data-mining.The paper discussed the attributively similarity of spatial entity,and carved it up two kinds:the similarity of attribute of name and semantic.Ueing for reference of the calculating method of list-similarity,brang out the calculating method based on the similarity of attribute of name,and applied in the distilling of spatial association rules.The method advanced in the universality and applicability.
Keyword: Similarity Of Attribute,Spatial Association Rule

引言

空间关联规则表达的是由一个区域单元上单个或一类地理现象（空间实体）与邻近地理区域单元上其他地理现象（空间实体）在空间（几何）或属性信息上的相关性而得出的知识性信息。空间关联规则提取是空间数据挖掘的重要组成部分同时也是基本任务之一。其目的在于发现空间实体间的相互作用、空间依存、因果或共生的模式。提取的依据包括空间拓扑关系、空间距离关系、空间方位关系、空间实体属性相似性等关联性指标。空间关联规则的

基金项目：国家 863 计划资助项目（2007AA12Z206）。

提取是一个复杂的过程，需要用到几何学、统计学、计算机视觉、模式识别、人工智能和专家系统等领域的理论和方法。

自 K. Koperski 在 1995 年将传统关联规则拓展到空间数据挖掘领域以来^[2]，很多学者对空间关联规则的概念、挖掘算法、不确定性的表达和挖掘结果的可视化等方面都进行了深入的研究并取得了一系列的成果^[1]。比较有代表性的包括袁春红（2004）提出了元规则指导下逐步求精的多层空间关联规则挖掘算法^[2]；何彬彬（2007）将空间统计分析应用于空间关联挖掘领域。这些研究均取得了一定的研究成果，但共同点是对基于实体间空间关系（拓扑、距离、方位关系）提取空间关联规则较多，而对基于实体间的属性相似性提取空间关联规则的研究较少。

1 属性相似性

属性相似性是指空间实体属性项之间在语义信息上相似。空间数据具有的属性特征决定了空间实体之间存在着某种属性上的关联。这种关联是某一项或几项属性项之间的相同或相似。基于属性相似性提取空间关联规则是通过实体的某一项或几项属性进行属性匹配，找出属性相同或相似的实体，再经过整理表达，从而提取不同实体间空间关联规则的过程，这在本质上这是一个属性匹配的过程。

目前，国内外很多学者在都从理论上做了比较深入的研究，并取得了很多的研究成果。

1998 年 Cobb^[3]研究了美国矢量数据标准格式 VPF 文件的合并问题，由于 VPF 文件具有丰富的属性信息 Cobb 提出了基于语义的属性匹配方法，具体方法是通过属性值的模糊相似度计算实体文字型属性值之间的相似性。例如在比较公路段“路面性质”属性项时，GPS 数据中属性值为“沥青”，而别的数据源中该属性项值可能为“硬”，虽然这两个值并不一致，但其语义基本一致。1999 年 Stock K. and Pullar D 提出了基于谓词逻辑表达式的匹配方法对不同实体的属性项进行相似性匹配；近年来又出现了通过模糊聚类进行属性匹配的方法，例如通过属性项中的频繁项来衡量相应属性项的匹配度，或采用属性项的名称、语义相似性对属性项进行匹配度计算。

综合比较各种匹配算法总体上来说和实际的应用要求还有一定距离。在实际应用中比较常用的是基于“编辑距离”的属性相似性匹配方法，即把属性项按照字典书序进行排序，根据插入、交换、删除和替换的操作数判断两者之间的距离。这种算法实现简单，但匹配准确率较低，而且不能解决异名同义、缩略词等问题。如给定三个属性名称 aaaabbbb、bbbaaaa 和 abcdefg，名称 1 到名称 2、名称 3 的编辑距离都是 6，但是实际上名称 1 与名称 2 之间的相似度要高得多^[4]。根据属性相似的程度不同，属性相似性可以分为属性名称相似性和属性语义相似性两种。相应的空间关联规则提取方法也分为基于属性名称相似性提取法和基于属性语义相似性提取法两种。

2 空间关联规则提取方法

2.1 基于属性名称相似性提取

名称是实体识别的重要特征，在较大地理范围内，同一实体的名称会存在差异。如福建

省的“后山-岭腰公路”另一小段又称为“岭腰-江上公路”，在另一段称为“江上-李屯洋公路”等。另外，命名规范的差异也会导致不同来源空间数据中的同一空间实体存在名称上的差异，如“同三线高速公路”在福建省的另一段称为“漳诏线”。这种差异增加了空间数据的复杂性，但同时也为建立空间实体间的关联关系提供了一定的条件。属性名称相似性不仅仅存在于不同数据源的同一空间实体之间，同时也存在于同一数据源的同类实体之间。如图 1（1）所示：在四个公路桥梁中，“院前桥”和“院前中桥”名称相似，“周厝 1#中桥”和“周厝 2#中桥”名称相似。在图 1（2）中，“大西山”和“小东山”的名称也具有很高的相似性。

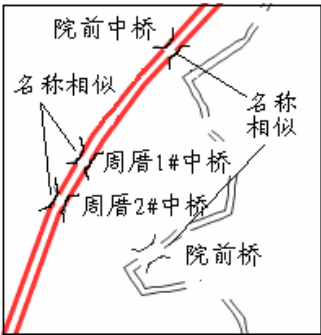


图 1 属性名称相似性（1）

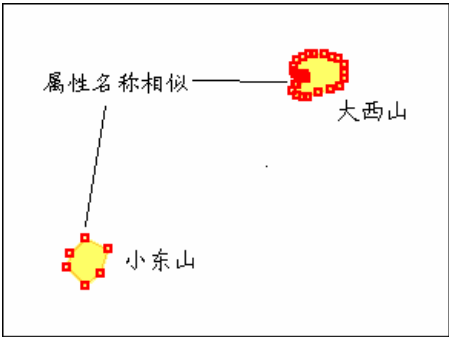


图 1 属性名称相似性（2）

基于属性名称相似性提取就是通过对实体的相关属性项进行名称匹配，查询记录与目标实体属性项内容相似或完全相同的实体集合，再经过整理表达，从而建立目标实体与参考实体间的关联联系的过程。基于属性名称相似性提取的关键是属性项名称相似性计算。在本文中借鉴文献^[5]中提出的序列之间相似度的计算方法，对属性名称相似度进行计算。

设 ξ 是一个符号序列，一个空间实体的属性项即是由 ξ 中的符号序列组成， ξ 中符号的个数即为该属性项的长度，两个属性项的相似度可以通过计算相对的条件概率分步进行计算。设实体 A 的属性项为 $S = s_1 \cdots s_m$ ，实体 B 的相应属性项为 $T = t_1 \cdots t_n$ ，设 $m < n$ ，则属性项 S 和属性项 T 之间的属性相似性可以定义为：

$$\text{sim}(S,T) = \frac{\prod_{i=1}^m p_T(s_i | s_1 \cdots s_{i-1})}{\prod_{i=1}^m p(s_i)} = \frac{\prod_{i=1}^m \frac{p_T(s_i | s_1 \cdots s_{i-1})}{p(s_i)}}{1}$$

$p_T(s_i | s_1 \cdots s_{i-1})$ 表示在属性项 T 中 s_i 正好是 $s_1 \cdots s_{i-1}$ 的下一个符号的条件概率。 $\prod_{i=1}^m p(s_i)$ 表示属性项 S 由无记忆随机产生器产生的概率， $p(s_i)$ 表示在属性项 S 中任何位置出现 s_i 的概率。

在上述定义中，属性项的匹配是从属性项的开头进行匹配的，但有时候不一定总从属性项的开头进行匹配，此时就要对上述定义进行修正与扩展。

定义属性项的名称相似度为：

$$N\text{sim}(S,T) = \frac{\sum_{1 \leq i \leq j \leq m} \text{sim}(s_i \cdots s_j, T) * (j - i + 1)}{\sum_{1 \leq i \leq j \leq m} (j - i + 1)}$$

称为属性项 S 和属性项 T 之间的属性名称相似度, 其中, 属性项 $S = s_1 \cdots s_m$, 属性项 $T = t_1 \cdots t_n$ 。同样, 在上述定义中设 $m < n$ 。经过修订后, 属性名称相似度在定义的广泛性、适用性方面都有了提高。

基于属性名称相似性提取是通过对空间实体两两属性项之间内容的名称相似性匹配, 从而提取空间实体间的关联规则, 是狭义的基于属性相似性提取。

2.2 基于属性语义相似性提取

基于属性名称相似性提取的方法简单实用, 算法实现效率也比较高, 但如果实体属性项中存在表述不同而语义一致或者存在缩略表述的情况, 则提取的成功率就会比较低, 例如“英吉利与北爱尔兰联合王国”与“英国”虽然其属性名称相似度比较低, 但实际上语义一致。

基于属性语义相似性提取就是通过对实体的相关属性项进行语义匹配, 查询记录与目标实体属性项内容语义相似或相同的实体集合, 再经过整理表达, 从而建立目标实体与参考实体间的关联联系的过程。基于属性语义相似性提取是广义的基于属性相似性提取。

与基于属性名称相似性提取相似, 基于属性语义相似性提取的关键是空间实体属性项之间的语义相似性判断与计算。目前, 很多学者在语义匹配方面做了很多的研究, 并取得了很多研究成果。张燕^[6]等提出了一种基于网格服务的语义匹配方法, 用一个语义数据库存储所有的网格服务的语义信息, 通过建立相关领域的本体库从而为服务功能的语义匹配提供支持。谢红薇^[7]等提出了一种通过概念图语义匹配的方法来计算两个实体间的属性语义相似性, 该方法把用户的提取要求转化成一个概念图, 然后通过和数据库中的概念图进行匹配计算实体间的语义相似性。张茅元^[8]等提出了义素相似度的定义, 进而提出了基于语义信息匹配的网页信息识别和提取方法。刘青宝^[9]等参照文献提出了语义相似度的计算方法, 并提出了基于模糊聚类的语义相似性计算方法, 并通过等价闭包法对实体属性进行模糊聚类, 从而得出实体属性间的多层次属性匹配结果。

由于矢量空间数据的复杂性, 以及矢量空间数据的不一致性, 语义匹配算法对数据模型以及属性数据类型有很大依赖, 由于不同的数据源具有不同的数据特征, 使得相应的语义匹配算法也有所不同。同样, 基于属性语义相似性提取也需要针对具体的数据源而选取相应不同的语义匹配算法。

3 实例应用

本实例的目的是探讨基于属性相似性提取在空间关联规则提取中的应用, 并对其有效性进行评估。本实例选择的数据源是福建省空间实体数据库, 采用的方法是基于属性名称相似性提取, 通过对实体相应属性项进行名称相似性判断计算, 提取出属性项相似的空间实体, 进而提取出实体间的空间关联规则。具体步骤如下。

首先通过数据选择和查询, 将福建省空间实体数据库中的油库类实体作为提取数据源, 并将油库的“岩、土质特性”作为目标提取属性项。

然后针对该属性项进行两两目标间的属性名称相似性匹配, 并将匹配结果记录在结果表集中。如表1所示。

表 1 基于属性名称相似性提取结果表集

基于属性名称相似性提取	属性名称相似度
<油库 1, 油库 2>	100%
<油库 1, 油库 3>	100%
...	...
<油库 1, 油库 14>	100%
<油库 2, 油库 3>	100%
...	...
<油库 13, 油库 14>	100%
总记录数（非重复）	91

由上表可得福建省空间实体数据库中油库间的空间关联规则，即：

$$is_ (X,油库) \rightarrow 属性名称相似性 (Y,沙质粘土) \quad (置信度: 100\%)$$

即福建省的油库均建设在沙质粘土类的土质之上，置信度 100%。

4 结论

空间关联规则提取是空间数据挖掘的重要内容，本文在分析空间实体属性相似性的基础上，将空间实体属性相似性分为属性名称相似性和属性语义相似性两种，借鉴序列相似性计算方法，给出了属性名称相似性计算方法，并将其应用在空间实体空间关联规则提取当中，该方法在提取方法的广泛性、适用性上都有一定的提高。

参考文献

[1] M Ester, H P Krlgel, J Sander. Spatial data mining: A database approach, in Advances in Spatial Databases, 1997, 1262: 47—66.

[2] 袁红春, 熊范伦. 元规则指导下的逐步求精多层空间关联规则挖掘算法[J]计算机工程, 2004, 30 (8): 34-36.

[3] Cobb M., Chung M., Foley H.. A Rule-based Approach for the Conflation of Attributed Vector Data[J], GeoInformatica, 1998, 2 (1):7-35.

[4] 刘青宝, 金燕, 邓苏等. 基于模糊聚类的属性匹配算法. 模糊系统与数学[J], 2006, 20 (6): 96-102.

[5] Yang J , W ang W. CLU SEQ: efficient and effective sequence clustering [A]. P roceedings of the 19th IEEE International Conference on Data Engineering (ICDE) [C]. 2003.

[6] 张燕, 王锋, 张睿. 基于本体的网格服务语义匹配方法[J], 计算机工程, 2007.4

[7] 谢红薇, 李瑞霞, 余雪丽, 于晓霞, 基于概念图匹配的语义相似性算法研究[J],软件时空, 2007 年第 23 卷第 7-3 期。

[8] 张茅元, 邹春燕, 卢正鼎, 一种基于予以匹配的 Web 信息提取方法研究[J],计算机工程与应用, 2006.23

[9] 周辉, 鲁燕飞, 王黔英, 袁芳, 基于信息粒度的属性权重确定方法[J],知识丛林

作者简介

王海涛（1982—），男，河南新乡人，博士研究生，主要研究领域是 GIS 应用开发和数字制图技术。

最快完成车辆路径问题的改进蚁群算法

闻思源¹ 魏红翠²

(1. 山东经济学院 信息管理学院, 山东省 济南市 250014;

2. 山东经济学院 图书馆, 山东省 济南市 250014)

摘 要: 本文考虑以改进蚁群算法, 解决完成时间最小为目标的车辆路径问题, 首先给出以完成时间最小为目标的车辆路径问题, 然后给出该问题的新的车辆分配方法和目标函数计算方法, 并给出一种局部搜索方法, 进而给出求解该问题的改进蚁群算法, 最后给出求解算例。

关键词: 车辆路径问题; 蚁群算法; 局部搜索; 最快完成时间

Modified Ant Colony Algorithm for Fastest Complete Vehicle Routing

WEN Si-yuan¹ WEI Hong-cui²

Abstract: Ant colony algorithm for time balancing vehicle routing problem is considered. First vehicle routing problem which objective is balancing the vehicle time utilization is given, then chromosomes method and computer method of objective function are given. Next ant colony algorithm for time balancing vehicle routing problem is given. At last computational instances are given.

Keywords: Vehicle routing problem; Ant colony algorithm; Local search method; Fasted Finished time

1 引言

车辆路径问题在交通和物流配送领域有着非常重要的应用, 因而是人们研究的热点问题, 并已经取得了很多研究成果[1]。车辆路径问题是个 NP-hard 问题[2], 因而人们主要考虑求解该问题的启发式算法或进化算法, 禁忌搜索算法、遗传算法、模拟退火算法和蚁群算法都被用来求解该问题, 有关工作的综述也比较多[3,4]。

车辆路径问题根据其问题不同可以分成很多类型, 人们研究比较多的有能力约束的车辆路径问题(CVRP)、带时间窗口的车辆路径问题(VRPTW)、多发点的车辆路径问题(MDVRP)、动态车辆路径问题(DVRP)等[4]。文[5]首次提出关于以平衡车辆时间效用为目标的车辆路径问题(记为 BTVRP), 并且提出一个有效的启发式算法。该问题在满足所有客户的需求前提下, 追求不同车辆时间效用差距最小的。张新、马建华[6]等在平衡时间效用车辆路径问题的基础上提出了以最长车辆完成时间最小为目标的车辆路径问题, 称之为最快完成车辆路径

问题（记为 FCVRP）。该类问题立足于用最短的时间完成所有的服务要求，主要出现在服务时间的重要性远大于服务费用的物资配送中，诸如应急物资发放、早报的配送、特快专递、快餐外卖等情况，因而该问题的研究具有现实意义。

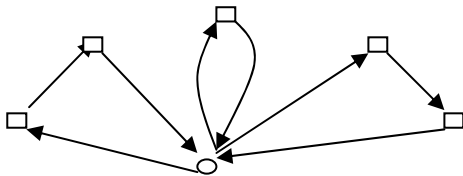
文[6]借助文[7]中的方法给出了求解完成时间车辆路径问题的蚁群算法，该算法从计算结果上看非常理想，但由于其在计算目标函数时多次计算最短路，因而计算速度比较慢。本文针对完成时间车辆路径问题提出新的编码方式和改进方法，并与文[9]种的方法进行比较。首先第 2 节给出完成时间车辆路径问题，然后第 3 节给出该问题新的目标函数值计算方法，第 4 节给出一个局部搜索方法，第 5 节给出改进后的蚁群算法，最后第 6 节给出计算实例和计算结果比较分析。

2 完成时间车辆路径问题

假设有一个发货点和 n 个顾客，分别用 $0,1,2,\dots,n$ 表示。每个顾客有一个大于 0 的需求量 $q_i, i=1,2,\dots,n$ ，发货点的货物总量多于顾客的需求总量。在发货点有 m 个车辆，每一个车辆的运输能力为 w 。顾客之间或顾客与发货点之间的运行时间为 $t_{ij}, i=0,1,\dots,n, j=0,1,\dots,n$ ，并且 $t_{ij} = t_{ji}$ 。车辆路径问题就是安排每辆车的服务顾客群以及服务顺序以满足以下要求：

- 1. 每个顾客由且只由一个车辆提供服务，设 $q_i \leq w, i=1,2,\dots,n$ 。
- 2. 每个车辆可以服务多个顾客，但其服务顾客的需求总量不超过其运载能力。
- 3. 每辆车从发货点出发依次经过所服务的顾客，最后返回发货点。

具体如下图所示。



其中 ○ 代表发货点，□ 代表顾客

图 2.1

一般的车辆路径问题是以所有车辆的总行程或者总运行时间最小为目标，这主要是从车辆运输成本上考虑问题，而对于顾客等待服务的时间并不是很在意。带有时间窗口的车辆路径问题虽然也考虑了顾客接收服务的时间因素，但只是要求每个顾客的接收服务的时间落在某个给定的区间，具有很大的弹性。在这些传统的车辆路径问题中时间并不是紧迫的因素，而在有些情况下时间会变得非常重要，等待的时间越长说明服务质量越差，顾客的满意度直接与其等待的时间成正比，因而从占有市场和开发顾客的角度讲，以最短的时间满足所有顾客的需求是非常重要的。

只有当每一个顾客都得到服务（或所需的物资）时整个任务才算结束，因而平衡时间的车辆路径问题不是以某一个顾客得到服务为标准，而是以最后一个得到服务的顾客的时间为标准。显然最后一个得到服务的顾客一定是在某个车辆线路的最后一个顾客，而且它的服务时间就是该车辆到达它所经过线路的使用时间。假设某个车辆经过的线路为 $0 \rightarrow i_1 \rightarrow i_2 \rightarrow \dots \rightarrow$

i_k ，则顾客 i_k 就是其最后到达的顾客，它的服务时间为

$$T_{i_k} = t_{0i_1} + t_{i_1i_2} + L + t_{i_{k-1}i_k}$$

也称其为该车辆的完成时间。

给定车辆安排其中完成时间最长的车辆称为关键车辆，该车辆的完成时间就是整个任务的完成时间。设每个车辆的完成时间分别为 $T_j, j=1,2,\dots,m$ ，整个服务完成的时间为：

$$T = \max_{j=1,2,\dots,m} T_j$$

该时间也称车辆安排的完成时间。

显然车辆越多完成时间越短，但是车辆个数有限，完成时间车辆路径问题就是在使用车辆数不超过 m 的条件下，追求服务完成时间最小。

平衡时间效用车辆路径问题只是保证了车辆时间效用的差距达到最小，可能会出现每个车辆的完成时间都很大而他们的差距却很少的情况，因而平衡时间效用车辆路径问题不能保证最长完成时间达到最小。

文[6]给出了一个求解完成时间车辆路径问题的蚁群算法，该文借鉴文[10]中的编码方式，首先给出一个客户的排列，然后把该排列顺序分割成若干段，每段对应着一个车辆。在具体分割时文[9]把问题转化成有 $n+1$ 个点的有向图，其中点 0 代表发货点，点 $i=1,2,\dots,n$ 代表客户 $P(i)$ ，如果顾客 $P(i+1)$ 、 $P(i+2)$ 、...、 $P(j)$ 的需求量之和小于等于车辆运输能力 w ，则在点 i 和点 j 之间添加一条弧，弧 (i,j) 的边长代表路径 $0 \rightarrow P(i+1) \rightarrow P(i+2) \rightarrow \dots \rightarrow P(j)$ 的完成时间。然后在有向图上求解一个从点 0 到点 n 的一个 m -弧最短路，最短路的每一条弧对应着一个车辆服务的客户，其行车顺序是按照给定排列中的顺序进行，这样就得到了一个关于该排列的一个最长完成时间最短的车辆安排。

而求 m -弧最短路则是通过不断修改弧长限制后求边最少路来实现的，因而在求目标函数时需要多次求边数最少的路。文[6]中方法虽然计算效果很好但计算速度比较慢，下面本文将试图给出新的车辆分配方法和目标函数计算方法以提高计算速度，并对两种方法进行比较。

3 车辆优化分配方法

下面考虑一种车辆完成时间平衡调整方法，对于给定的客户排列，首先给出一个初始可行的车辆划分方法，然后根据车辆完成时间进行调整，目的是在不改变客户顺序的前提下把完成时间最长的车辆的完成时间降下来。

3.1 初始划分方法

按照客户的排列顺序依次给客户指定车辆，车辆沿着客户在排列中出现的先后顺序进行配送。在能力限制下优先使用以安排的车辆，只有在前一个车辆已经装满时才使用下一个车辆，具体步骤如下：

Step1 令 $i=1, l=1, b=0$ ；

Step2 如果 $b+q(i)$ 小于等于 w ，客户 i 安排第 l 辆车， $b=b+q(i)$ ， $i=i+1$ ，转 step3；

否则转 step4；

Step3 如果 $i > n$ 停止，安排完毕；否则转 step2；

Step4 如果 $l < m$ ， $l=l+1$ ，转 step2；否则停止没有可行安排。

3.2 优化调整方法

对于给定的划分方案，首先计算出每个车辆的完成时间，找出完成时间最长的车辆，然后试图减少该车辆的完成时间。在不改变客户排列顺序的前提下，能够减少该车辆的完成时间的方法有两个：

- 前向转移：把该车辆的第一个客户转给前面的车辆；
- 后向转移：把该车辆的最后一个客户转给后面的车辆；

这两种方法类似，下面以后向转移方法说明在什么情况下可以实现这种转移而使最长完成时间严格减少。

假设该车辆安排的关键车辆为 c ，其完成时间为 T_c ，服务的最后一个客户为 k ，而第 $c+1$ 辆车服务的客户序列为 i_1, i_2, \dots, i_l ，在不改变客户排列顺序的前提下只能把 k 放在该序列之前，得到一个新的序列 k, i_1, i_2, \dots, i_l 。如果该序列的完成时间小于 T_c ，则直接把 k 转移到第 $c+1$ 辆车得到一个新的划分，且新的划分的最长完成时间严格减少。否则考虑把新序列后面的客户再转移到第 $c+2$ 辆车中，如果转移最后一个客户后第 $c+1$ 辆车的新客户序列 $k, i_1, i_2, \dots, i_{l-1}$ 的完成时间严格小于 T_c 就转移一个，否则考虑转移多个，目的是转移最少的客户使得转移后的完成时间严格小于 T_c 。假设需要转移 r 个客户，第 $c+2$ 辆车原服务客户序列为 j_1, j_2, \dots, j_h ，则转移后得到的新序列为 $i_{l-r+1}, \dots, i_l, j_1, j_2, \dots, j_h$ 。如果该序列的完成时间小于 T_c ，则把 k 转移到第 $c+1$ 辆车，把第 $c+1$ 辆车最后 r 个客户依次转移到第 $c+2$ 辆车上，其他客户不变，即第 $c+1$ 和第 $c+2$ 辆车的新客户序列为

$$\begin{matrix} k, i_1, i_2, \dots, i_{l-r} \\ i_{l-r+1}, \dots, i_l, j_1, j_2, \dots, j_h \end{matrix}$$

如果第 $c+1$ 辆车新客户序列的完成时间不小于 T_c ，依照上述方法继续往下转移直至后面没有车辆为止。

- 这个过程最后必然出现以下两种情况之一：
1. 某个车辆接受上个车辆转移客户后的完成时间小于 T_c ；
 2. 转移到第 m 辆车后完成时间依然不小于 T_c 。

对于第一种情况可以实现转移，且得到的新最长完成时间严格减少。而对于第二种情况则不能进行转移，此时考虑前向转移方法。

经过最优调整最后得到的车辆安排称为顾客序列的最优安排，对应的完成时间称为顾客序列的完成时间。

考虑到我们产生初始车辆分配时先安排前面车辆，后面的车辆一般会有较多剩余能力，所以在实际计算时我们只采用了后向转移方法，而且每次最多转移一个客户。我们针对文[9]中的算例，随机产生 200 个排列，然后分别用弧最短路分割法和平衡调整法计算其最长完成时间。我们用两种方法计算的最长完成时间的差的绝对值占弧最短路分割法计算的最长完成时间的比例来表示两种方法结果的差距，具体计算结果如图 3.1 所示。

从中可以看出对于 91.5%以上的排列，两种方法的结果是一样的，对于不一样的排列两者结果的差距也没有超过 8%，平均差距为 0.25127%。在同样的计算环境下，第一种方法共使用了 0.5 分钟，而第二种方法共使用了 0.05 分钟，是第一种方法使用时间的十分之一，因此第二种方法在计算速度上明显优于第一种方法。

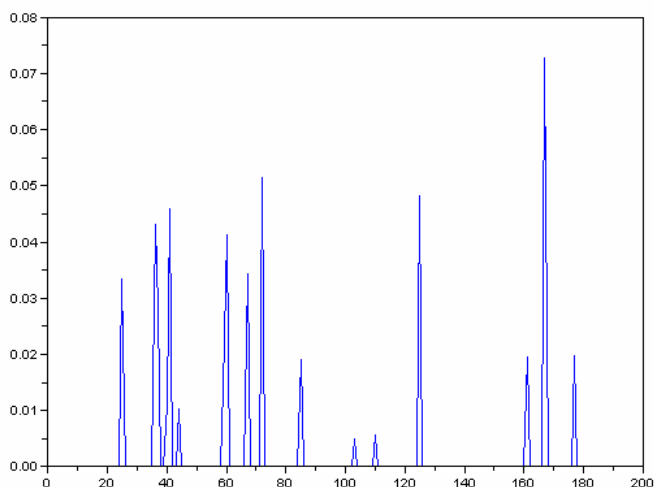


图 3.1 两种方法的计算结果差距

4 局部搜索算子

为了进一步改进计算结果，在产生一个编码后，我们引入了局部搜索算法，通过某一个或两个客户的位置变换得到一个新的排列，变换的目的是试图降低最长完成时间。

要减少一个车辆安排的最长完成时间，就必须减少关键车辆的完成时间。减少关键车辆的完成时间最直接的方法是从关键车辆顾客中拿走一个客户，客户从该车辆去掉后就必须加入新的车辆序列中，从而增加了该车辆的完成时间，因此我们把该客户放在完成时间最少的车辆中。假设关键车辆的客户序列为：

$$i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_l \quad (4.1)$$

完成时间最短的车辆的客户序列为

$$j_1 \rightarrow j_2 \rightarrow \dots \rightarrow j_h \quad (4.2)$$

其完成时间记分别为 T_1 和 T_2 。每一个客户从原序列拿出后完成时间都会改变，这个时间改变量称为客户的移出时间，记为 TO 。当距离或时间满足三角不等式关系时，移出时间必然为负值。

同样对于某个客户插入序列的某个位置后，序列完成时间发生变化，不同的插入位置改变量不一样，最小的改变量称为客户的移入时间，对应的位置称为最优位置。

分别计算出关键车辆的每个客户对于序列(4.1)的移出时间 $TO_k, k=1,2,\dots,l$ 和对序列(4.2)的移入时间 $TI_k, k=1,2,\dots,l$ 。

对每一个单移客户计算移动后两个车辆完成时间最长者，称为该客户的移动时间，记为 TY ，即：

$$TY_{i_k} = \max \{T_1 + TO_{i_k}, T_2 + TI_{i_k}\}, \quad k=1,2,\dots,l$$

移动时间最小的客户称为最佳移动客户，把最佳移动客户移到其最优位置后就可以得到一个新的排列，把该排列作为局部搜索的结果，利用上节的方法计算其最优完成时间。如果其最优完成时间小于原排列的最优完成时间，则替代原排列，否则以一定的概率替代原排列。

5 改进蚁群算法

在文[9]中的蚁群算法中使用本文的目标函数计算方法和局部搜索算子，排列生成方式和信息素处理方式依然按原算法中处理，就可以得到一个改进的蚁群算法。

具体步骤如下：

Step1（初始化）指定蚁群规模 n 、父代种群规模 m 、概率 p 以及 α, β ，输入初始信息素 B ， $t=0$ ，计算出权重矩阵 C ，令 $y_{ij} = c_{ij}^{\alpha} b_{ij}^{\beta}, i=1,2,\dots,n+1, j=1,2,\dots,n$ 。并根据 $Y = (y_{ij})_{n+1,n}$ 随机产生 n 个排列，记为：

$$A(0) = \{A_1(0), A_2(0), \dots, A_n(0)\}$$

Step2（计算完成时间）计算出每个排列的服务完成时间，并找出当前排列中完成时间最小的排列，判断是否满足要求，如果满足要求就停止，否则转向 Step3

Step3（局部搜索）对当前最优排列进行局部搜索产生一个新的排列，如果新排列的完成时间小于原排列的完成时间，就用新排列替代原排列；否则产生一个随机数 r ，如果 $r \leq p$ ，则新排列替代原排列。转 Step4。

Step4（选择）根据完成时间按照轮盘赌的方式随机选择 m 个排列作为父代种群。转 Step5。

Step5（计算新的信息量）令

$$\Delta b_{ij}^l(t) = \begin{cases} 1/v_l(t); \text{第} l \text{ 个排列中第} i \text{ 个数后紧随第} j \text{ 个数} \\ 0; \text{否则} \end{cases}$$

$$\Delta b_{ij}(t) = \sum_{l=1}^m \Delta b_{ij}^l(t),$$

$$b_{ij}(t+1) = \rho b_{ij}(t) + \Delta b_{ij}(t)$$

Step5（产生新蚁群）令 $y_{ij} = c_{ij}^{\alpha} b_{ij}^{\beta}, i=1,2,\dots,n+1, j=1,2,\dots,n$ ，并根据 $Y = (y_{ij})_{n+1,n}$ 随机产生 n 个排列，记为

$$A(t+1) = \{A_1(t+1), A_2(t+1), \dots, A_n(t+1)\}$$

令 $t=t+1$ ，转 step2。

6 计算结果分析

我们使用 Scilab 编写了上述算法的实现程序，设定种群规模为 8，父代种群规模为 6、 $p=0.1$ 以及 $\alpha=1.2, \beta=0.7$ ，迭代步数为 200。我们用该程序对文[6]中的算例进行计算，经过多次计算比较，发现计算结果与原算法的结果一样，而在同样的计算环境下改进算法使用的计算时间是原算法的三分之一，从而说明该算法在计算时间上的优势。

同时我们考虑算例中每段道路的时间发生变化对计算的影响，针对不同的车辆数分别随机产生 20 个算例，每段道路的通行时间在 9-20 分钟之间。然后用两种算法分别计算每一个算例，他们的完成时间的差距最大不超过 5%。并且多数情况下改进蚁群算法的结果优于或等于原算法的结果，对于不同车辆数的具体情况如表 6.1 所示。

表 6.1 结果比较

车辆数	3	4	5	6
改进算法结果变好的算例数	9	8	8	6
两种算法结果相同的算例数	7	7	8	12

从而说明在计算结果上改进算法也比原算法好。

参考文献

[1] C. Bodin, Twenty years of routing and scheduling, Operation Research 38 (1990) 571–579.

[2] J.K. Lenstra, A. Rinnooy Kan, Complexity of vehicle routing and scheduling problems, Networks 11 (1981) 221–227.

[3] Cordeau J-F, Gendreau M, Laporte G, Potvin J-Y, Semet F., A guide to vehicle routing heuristics. Journal of the Operational Research Society 2002;53:512–22.

[4] 刘云忠, 宣惠玉, 车辆路径问题的模型及算法研究综述, 管理工程学报, 19 (1): 124-130。

[5] G. Nikolakopoulou, S. Kortesis, A. Synefaki, R. Kalfakakou, Solving a vehicle routing problem by balancing the vehicles time utilization, European Journal of Operational Research 152 (2004) 520–527.

[6] ZHANG Xin, Ma Jianhua, Liu Weilong, Jin Fang, Ant colony algorithm for vehicle routing problem with shortest completion time, The proceedings of the 13th international conference on industrial engineering and engineering management, 2928-2933, 2006.8.

[7] Christian Prins, A simple and effective evolutionary algorithm for the vehicle routing problem, Computers & Operations Research, 31(2004):1985-2002.

作者简介

闻思源 (SiYuan Wen) 男, 1970 年出生, 硕士, 籍贯辽宁省沈阳市, 讲师, 主要研究领域为通信与信息系统, 最优化理论, 近年来, 已获得省部级科技进步奖 4 项, 在各级刊物和学术会议上发表论文 10 余篇, 其中 4 篇列入 EI、ISTP 检索。

基于训练文本特征扩展的中文短文本分类研究

闫 涛¹ 王细薇¹ 樊战伟²

(1. 河南城建学院 信息中心, 河南平顶山 467044;
2. 平顶山市平东热电有限公司 河南平顶山 467044)

摘 要: 针对短文本所描述信号弱的特点, 提出了一种基于训练文本特征扩展的中文短文本分类方法, 该方法在没有引入新特征的前提下, 在训练阶段用一种基于共现关系的特征权重调整方法实现训练文本特征扩展, 提高分类器自身的性能。实验证明, 这种方法具有高的分类性能, 其微平均和宏平均值都高于常规的文本分类方法。

关键词: 短文本分类; 训练文本; 特征扩展

A Method for Chinese Short Text Classification Based on Training set Feature Extension

YAN Tao¹ WANG Xi-wei¹ FAN Zhan-wei²

(1.Henan University of Urban Construction ,Network Information Center, Pingdingshan,467044,China; 2.Ping Dong Thermal Power Co.,Ltd,Pingdingshan, 467044,China)

Abstract: In this paper, based on the characteristics that short texts describe weak signals, A method for chinese short-text classification based on training set feature extension is proposed. In this method, to achieve the expansion of the train text by using a method which based on co-occurrence relationship to adjust the weight of characteristics, without new features.Experimental results show the proposed method performs well both of its Micro-F1 and the Macro-F1 are higher than those of conventional approaches .

Keywords: Short Text Classification; Training Set; Features Extension.

引言

现行的手机短信息、QQ 聊天、网友评论等短文本特征词语由于长度短、所描述概念信号弱, 单纯的基于词特征选择容易使文本表达主题分散, 特征词易被赋予较低权重, 影响分类器性能, 基于共现关系的特征扩展能很好地解决文本表达主题分散的问题^{[1],[2]}。

本文提出了一种基于共现关系的词条权重提升的计算方法, 用提升特征共现词的权重的方法实现训练文本特征扩展, 对文本的表达方式进行重新组织和替换, 解决短文本所描述概念的信号弱固有缺陷。

基于训练文本特征扩展的短文本分类方法，先根据结合语言知识和统计信息的扩展词表构造方^{[3],[4]}，利用关联规则挖掘算法抽取训练集中的共现词对^[5]，考虑特征之间的共现关系对特征权重的影响并根据共现关系形成概念特征，提高分类器自身的性能。

1 基于训练文本特征扩展的策略

现行的特征集选择包括词特征选择，字特征选择和概念特征选择等^{[6],[7]}，其中，先行的实际分类系统主要以选择词特征为主，有的在特定领域加入一些人工规则等。但是，由于词语本身存在对短语和上下文的依赖等现象，因此，单纯基于词形的技术中，把意义可能密切相关的词孤立提取，忽略了词语的语言学特征和相互关系，因此导致这种特征提取存在较大的局限性。

关联规则就是从企业历史海量数据中挖掘出描述数据项之间相互联系的有价值的知识。随着收集和存储在数据库中的数据规模越来越大，人们逐渐对从这些数据中挖掘相应的关联知识越来越有兴趣。挖掘关联规则的应用来源一般认为是市场购物分析。其中：美国某一家大型商业企业“啤酒加尿布”的故事就是对其最好的诠释^[8]。在商场销售中，根据每个小票的内容记录数据而发现的商品之间所存在的关联知识无疑将会指导商家的有关人员分析顾客的购买习惯。比如：顾客在购买牛奶时，是否也同时购买面包或会购买那个品牌的面包，显然能够回答这些问题的有关信息肯定会有效地指导商家进行有针对性的促销，以及进行合适的货架商品摆放。市场购物分析可以帮助超市主管很好地确定哪些物品可进行捆绑销售。

本文将这种思想用于短文本特征扩展，考虑共现关系对特征权重的影响。我们认为，强关联规则的前项一旦在文本中出现，后项必以一定概率出现，具体在训练文本特征扩展中，对于训练文本中的特征词先根据特征共现集进行共现词的扩展，由于该共现词本身就在训练文本中，因此只需根据共现关系调整该词的权重即可。

1.1 基于训练文本特征扩展的基本流程

训练文本特征扩展是指在训练阶段，对于训练文本中的特征词先根据特征共现集^{[5],[9],[10]}把该词对应的共现词的权重提高形成概念特征词，并没有加入新的特征，然后再进行特征选择。该方法与测试文本特征扩展方法基于相同的扩展词表^[5]，均是了解决短文本所描述概念信号弱的问题，但是应用对象和扩展方法不同。

基于训练文本特征扩展的流程如图 1.1 所示。

特征共现集的创建方法同^[5]，这里不再重述。

1.2 基于共现关系的特征权重提升计算公式

基于词特征选择的分类系统往往采用单纯的词频作为特征权重，往往导致重要的词语权重较低。基于训练文本特征扩展的分类方法中，对于某一共现词对，假设其前项出现，后项以某一概率也会出现，那么我们就根据共现词对前项出现的次数和共现概率提升后项的出现次数，突出了共现词对之间的相互影响对特征权重及分类的贡献。

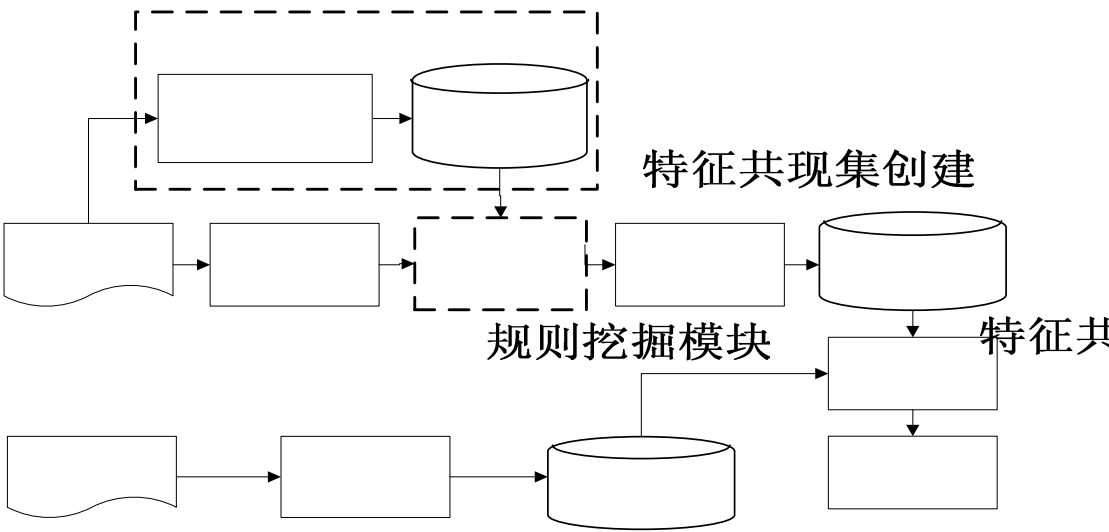


图 1.1 基于训练文本特征扩展的流程图

对于某一共现词对 $t_i \rightarrow t_j$ 后项特征权重提升计算

预处理

$$Wt_j = Wt_j \cdot \left(1 + \frac{Wt_i \times S \times C}{Wt_j} \right) \quad (1.1)$$

训练特征

其中， Wt_i 为特征 t_i 的权重， Wt_j 为特征 t_j 的权重， S 为共现词对 $t_i \rightarrow t_j$ 的支持度， C 为其置信度。本公式考虑了词频、共现关系支持度和置信度对特征权重的影响，引入语义概念信息，提升具有共现关系的特征权重，解决了单纯的基于词频的局限性。特征权重提升计算例子如图 1.2 所示。

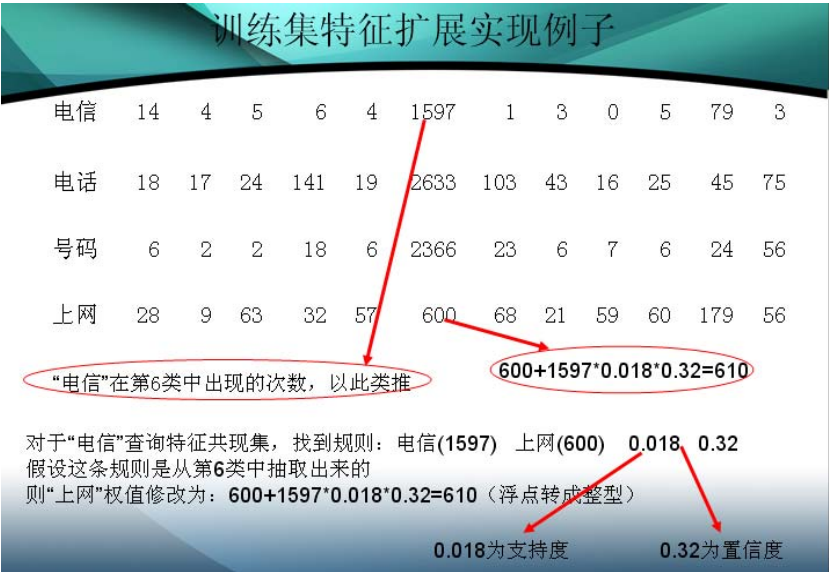


图 1.2 权值提升计算例子

1.3 训练文本特征扩展算法

以第 k 类为例($1 \leq k \leq 12$)

输入: 特征共现集 I_k ;

关联规则抽取阈值: 最小置信度阈值 C ;

最小支持度阈值 S ;

训练文本词频统计文件 `star.txt`;

输出: 训练文本词频统计结果 `star_.txt`;

步骤 1 对于 `star.txt` 文档中任一个特征词 t_i , 查询特征共现集 I_k , 如果存在唯一一个共现词对 $t_i \rightarrow t_j$, 并且当 C 大于设定的阈值时, 执行步骤 2。如果共现词对不唯一, 则计算 $S * C *$ (共现词对后项在文档中出现的次数), 则按照 $S * C *$ (共现词对后项在文档中出现的次数) 最大值的共现词对进行概念扩展, 执行步骤 2; 如果不存在共现词对则执行步骤 3;

步骤 2 根据特征权重提升计算公式 1.1 修改共现词对后项 t_j 的权值;

步骤 3 对特征词 t_i 不进行概念扩展。

2 实验和结果分析

2.1 实验数据

本章所使用的数据集是本项目组收集的共 12 个不同领域的共 470252 篇网友评论。其中财经类 35104 篇, 房地产类 28744 篇, 国际新闻类 42424 篇, 国内新闻类 48288 篇, 军事类 49320 篇, 科技类 37044 篇, 女性类 36032 篇, 汽车类 40372 篇, 书评类 39440 篇, 体育类 38512 篇。游戏类 38660 篇, 娱乐类 36312 篇。将每类文本集随机地平均分为四份, 以其中一份构成测试集, 另外三份构成训练集。本章的实验系统是在 WindowsXP 系统下, 使用 Borland C++ Builder 6.0 作为开发工具。

对文本分类的性能采用如下四种指标进行评估: 宏平均(Macro-F1)、微平均精确率(Micro-P)、微平均召回率(Micro-R)、微平均 F1 值(Micro-F1)。

2.2 实验方法设置

(1) 常规方法: 选用了清华大学开发的 CsegTag3.0 对中文进行分词, 并去除停用词, 然后用词频(Term Frequency)进行特征抽取, 采用 CHI 选择特征方法, 以朴素贝叶斯(Naïve)为分类器^[11], 训练集特征未经过概念扩展, 分别选取特征数为 1000、2000、3000.....10000, 进行 10 轮 12 类短文本分类实验。

(2) 基于训练文本特征扩展方法: 选用清华大学开发的 CsegTag3.0 对中文进行分词, 并去除停用词, 然后用词频(Term Frequency)进行特征抽取, 采用 CHI 选择特征方法, 以朴素贝叶斯(Naïve)为分类器^[11], 训练集特征经过概念扩展(特征共现集的创建^[5]中的方法, 不再重述), 分别选取特征数为 1000、2000、3000.....10000, 进行 10 轮 12 类短文本分类实验。其中, FP-Growth 的支持度和置信度阈值分别设为 0.3%、2%。

2.3 实验结果分析

(1) 特征数目不同时两种方法的分类性能比较。结果如图 2.1 所示。

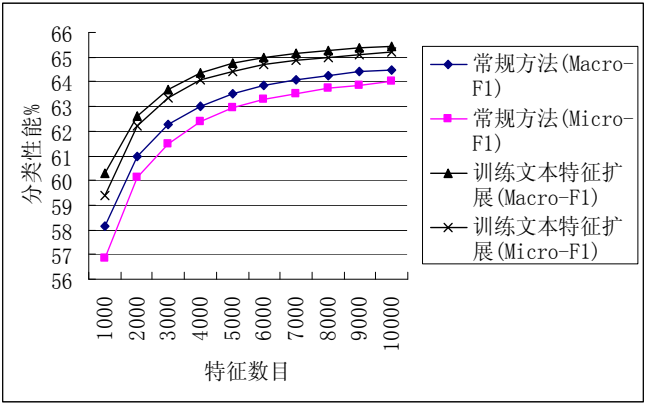


图 2.1 特征数目不同时两种方法的分类性能比较

上图的实验结果表明：总体上来说训练文本特征扩展方法比常规方法分类性能提高，前者比后者宏平均 F1 值提高 1-2 个百分点，微平均 F1 值提高 1-3 个百分点，基于训练文本特征扩展方法把特征词对应的共现词的权重提高形成概念特征词，有效的选择出了高性能的特征，提高了分类器性能，在一定程度上解决了短文本有效特征不足所造成的分类器分类性能下降的问题。

(2) 不同文档类别的性能分析

由 2.1 图可知，多类短文本分类性能在特征数目为 6000 时趋于稳定，上表是特征数目为 6000 时不同类别文档的分类性能比较。

表 2.1 特征数目为 6000 时不同类别的性能分析

性能 类别	常规方法			训练文本特征扩展方法		
	精确率%	召回率%	F1%	精确率%	召回率%	F1%
第 1 类	70.47	62.59	66.3	74.07	63.89	68.61
第 2 类	72.59	69.79	71.17	68.99	74.97	71.85
第 3 类	47.53	46.26	46.88	63.54	42.67	51.05
第 4 类	59.12	55.75	57.39	63.60	56.96	60.10
第 5 类	58.03	61.14	59.55	56.62	67.41	61.55
第 6 类	73.54	57.49	64.53	45.40	67.97	54.43
第 7 类	58.48	65.23	61.67	55.40	71.42	62.40
第 8 类	76.05	72.06	74.00	82.08	71.64	76.50
第 9 类	76.84	81.26	78.99	67.86	85.37	75.61
第 10 类	69.52	65.78	67.60	69.56	68.76	69.15
第 11 类	64.64	64.26	64.45	84.78	57.63	68.62
第 12 类	46.75	62.67	53.55	71.42	51.77	60.03
宏平均 F1%	63.84			64.99		
微平均 F1%	63.29			64.72		

由表 2.1 同样可以看出：基于训练文本特征扩展方法比常规方法在特征数目均为 6000 时的分类性能得到了提高，宏平均 F1 和微平均 F1 值都提高了约 1 个百分点。以精确率为例，可以看出，两种方法的精确率变化是比较大的，在-28%-25%之间波动，第 6 类科技类是降低了 28%，主要原因是这个类中抽出的共现词对中很大部分是在意义上没有多大关联的词，比如“电话”与“经历”，“碰到”与“详细”，这些词对虽然经常出现，但是并不是很好的特征。对比提高幅度最大的第 12 类 25%，虽然抽出的共现词对不多，但是抽出来的是关联意义很大且特征描述能力比较强的共现词对，比如“玩家”与“游戏”，“楼”与“主”，这与每个类别的语料信息有关。因此，需要设计适合中文短文本数据的关联规则挖掘算法，这有待于进一步的研究。

3 结语

本章利用一种基于共现关系的特征权重调整方法来实现对训练文本的特征扩展，在 12 类共 470252 篇网友评论构成的语料集上进行多组分类对比实验，性能和准确度均超过了常规分类方法，实验证明，基于训练文本特征扩展的短文本分类方法在理论和大规模数据实践上是可行和有效的，利用这种方法可以训练出好的分类器，在一定程度上解决了训练文本特征所描述概念信号弱、有效特征不足的问题，但是分类效果还不理想，因为分类阶段的测试文本特征本身长度短、所描述概念信号弱的固有缺陷没有得到改善。下一步的工作会考虑在扩展训练文本特征的基础上扩展测试文本特征，真正的改善短文本所描述概念信号弱的问题。

同时，以目前的实验结果来看，支持度、置信度、词频三个参数对分类性能的影响程度都能重要，但是，至于谁的重要程度更大，需要设计一个合理的特征权重调整公式来衡量支持度、置信度、词频对分类性能的影响，本章只是做了一个初步探讨，期待下一步的工作。

参考文献

- [1] Zelikovitz, S and Marquez, F. Transductive Learning for Short-Text Classification Problems using Latent Semantic Indexing[J]. International Journal of Pattern Recognition and Artificial Intelligence, Vol.19(2), 143-163, 2005.
- [2] Zelikovitz, s. Transductive LSI for Short Text Classification Problems[C]. In: Proceedings of the 17th International FLAIRS Conference, 556-561, 2004.
- [3] Lin D, Pantel P. Concept Discovery From Text[C]. Proceedings of Conference on Computational Linguistics 2002. Taipei, Taiwan. 2002: pp.577-583.
- [4] 张映海. 基于概念的中文文本检索研究[D]. 重庆大学. 2007, 4(30).
- [5] 王细薇, 樊兴华, 赵军. 一种基于特征扩展的中文短文本分类方法. 计算机应用[J]. 2009, 29(3): 843-845.
- [6] 王元珍, 廖莎莎, 江铭虎. 中文文本分类中基于概念屏蔽层的特征提取方法[J]. 中文信息学报, 2005, 20(3): 0022-0028.
- [7] 赵鹏, 耿焕同, 蔡庆生. 一种基于语义和统计特征的文本特征表示方法[J]. 小型微型计算机系统, 2007, 7(7), 1311-1313.
- [8] 张会容. 关联规则挖掘的研究与及其应用[D]. 华南理工大学, 2004-12-15.

- [9] 王元珍, 钱铁云, 冯小年. 基于关联规则挖掘的中文文本自动分类[J]. 小型微型计算机系统, 2005, 26(8): 1380-1383.
- [10] 中文停用词表.<http://download.csdn.net/source>.
- [11] 周茜, 赵明生, 扈雯. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 17—23.

作者简介

闫涛 (1970—), 男 (汉族), 河南洛阳人, 硕士, 河南城建学院副教授, 主要研究方向: 网络信息处理、网络安全 (wangxw@hncj.edu.cn);

王细薇 (1982—), 女 (汉族), 河南许昌人, 硕士, 主要研究方向: 中文信息处理。

樊战伟 (1971—), 男, 河南平顶山人, 平顶山平东热电公司高级工程师, 主要研究方向: 电力信息化。

基于Jacobi旋转的稀疏矩阵对角优势强化方法

银福康 宋君强 吴建平

(国防科学技术大学计算机学院, 长沙 410073)

摘要: 提出了一种基于 Jacobi 旋转的稀疏矩阵对角优势强化方法, 该方法尽可能在具有相同非零元结构的行列中寻找主元, 以保持矩阵的稀疏性。同时, 在某一行已经对角占优的情况下, 选主元时不考虑该行的元素, 该方法减少了对对角占优性无影响的迭代次数。实验结果表明, 相对于国际上现有的算法, 该方法能够有效减少迭代次数和过程填入元, 并且最终得到的结果矩阵中非零元个数明显减少。

关键词: Jacobi 旋转; 对角占优; 填入元; 稀疏矩阵

A Method for Strengthening Diagonally Dominance Based on Jacobi Rotation in Sparse Matrix *

YIN Fu-kang SONG Jun-qiang WU Jian-ping

(School of Computer, National University of Defense Technology, Changsha, 410073, China)

Abstract: A method for strengthening diagonally dominance based on Jacobi Rotation is proposed for general sparse matrices. In order to maintain the sparsity of the matrix, the pivot is selected as far as possible in those rows and columns with same non-zero structure. Meanwhile, the pivot selection does not consider the rows already diagonally dominant. This method can reduce the number of iterations which do not influence the diagonally dominance. Results of the experiment have shown that, compared to the available algorithms in the world, the provided method can reduce the number of iterations and fill-ins in progress, and the number of nonzero elements in the final matrix.

Keywords: Jacobi Rotation; Diagonally Dominance; Fill-Ins; Sparse Matrix

1 简介

矩阵固有属性的充分利用与挖掘在线性方程组的求解中具有重要意义, 如果系数矩阵对角占优, 则求解方法更有效, 且对方法的选取也不敏感。在基于 LU 分解的直接求解方法中, 当系数矩阵对角占优时, 其稳定性更好, 解的精度也越高。对迭代法, 在系数矩阵对角占优时, 即使最简单的 Jacobi 迭代与 Gauss-Seidel 迭代也能确保其收敛性, 且对角占优性越好, 收

受国家自然科学基金项目(60803039)资助。

敛速度越快。在基于 Krylov 子空间的投影型迭代法中,收敛速度依赖于系数矩阵的特征值分布,分布越集中,收敛越快,当系数矩阵具有对角占优性时,其特征值分布也相对更为集中,从而收敛速度必然很快。同时,在构造迭代法中的预条件时,如果相应的系数矩阵对角占优,则很容易构造更为有效的预条件。此外,在代数多重网格迭代法中,如果系数矩阵对角占优,则粗网格的构造相对简单,且所得到的多重网格型迭代的收敛速度也更快^[1]。

由于对角占优性在线性方程组求解中的重要性,国内外有大量学者对对角优势强化技术进行研究。Olschowka 与 Neumaier^[2]、Duff 与 Koster^[3]先后考虑采用非对称置换来改善非对称矩阵的对角优势,吴建平等针对对称矩阵,考虑了利用图分割技术与对称置换相结合来将绝对值较大的元素尽量交换到靠近对角线的位置,以一定程度上改善块对角优势^[4]。同时,吴建平在其博士论文^[5]中提及利用 Jacobi 变换等简单变换增强矩阵对角优势的思想,Jin Yun Yuan 和 Plamen Y. Yalamov 最近提出了一种基于 Jacobi 旋转进行对角优势强化的方法,并进行了具体算法设计和实现^[6]。

本文针对 Jin Yun Yuan 和 Plamen Y. Yalamov 所提出的方法进行了改进,在原方法中,每次迭代选取矩阵中非对角线上绝对值最大的元素作为主元来构造旋转矩阵。该算法没有考虑到选主元的次序对算法执行效率的影响。我们分析发现,不同的主元选取次序对迭代次数、填入元个数和最终结果矩阵中的非零元个数有明显影响。在稀疏线性方程组的求解中,较少的填入元能减少内存需求,还能减少乘除法次数^[7]。因此,保持矩阵的稀疏性很重要。同时我们还发现当某一行已经对角占优时,选主元时不考虑该行的情况下能够有效减少迭代次数。实验结果表明,与原算法相比,本文提出的算法在迭代次数、填入元、最终结果矩阵中非零元个数方面都得到明显改进,尤其是迭代次数,减少更为明显。

2 基于Jacobi旋转构造对角占优预处理器的方法的分析

Jin Yun Yuan 和 Plamen Y. Yalamov 于 2006 年提出了一种基于 Jacobi 旋转的对角优势强化方法^[6],该算法主要通过选取最大元作为主元,然后利用奇异值分解构造旋转矩阵进行行列变换。该算法选取最大元作为主元,简单实用。对于阶数不大的矩阵,能够以不超过矩阵阶数的平方次迭代下达到对角占优性。对病态矩阵,变换后矩阵的条件数能够变好。但是,该算法没有考虑到不同的主元选取策略对矩阵稀疏性、和变换过程中增加的非零元的影响。当矩阵是稀疏矩阵时,保持矩阵的稀疏性对于节省存储空间和减少乘除法次数影响较大。在原来的算法中,当某一行已经对角占优时,下一次选主元仍还可能选到该行的非零元,导致迭代次数无谓增多。特别是当各行中的元素幅度相差较大时,尤为明显。

例如:对矩阵

$$A=\begin{bmatrix}2.0e+010 & 2.0e+003 & 1.0e+005 \\ 600 & 800 & 400 \\ 0 & 1000 & 900\end{bmatrix}$$

如果取对角占优程度评判条件为: $\text{abs}(A(i,i))/\text{sum}[i]>=0.501$ ($\text{sum}[i]$ 为矩阵 A 的第 i 行元素绝对值之和)。显然第一行已经占优了,但是选主元时,会选择 $1.0e+005$,此次迭代对矩阵的对角占优性并没有影响。如果选择 1000,则可以增加一行的对角占优性,从而有利于减少总的迭代次数。

3 一种新的主元选取策略

基于第2节中的考虑,可以在执行 Jin Yun Yuan 和 Plamen Y. Yalamov 给出的算法过程中,主元不在对角占优的行中选取,即只在没有满足所需要对角占优程度判定条件的行中寻找主元。一般而言,可以预计这有利于减少迭代次数,但是从第4节的实验结果可见,最终所得矩阵中的非零元个数较原算法有时会增加,与我们预期的非零元减少相冲突。而且该方法仍然没有考虑矩阵的非零元分布与填入元的关系。为此这里继续在该方法的基础上,进一步利用结构相似性的概念来改进主元选取策略,以减少填入元的产生。

3.1 结构相似矩阵概念

设 A 是 $n \times n$ 实矩阵, MF 为 A 的结构相似矩阵。 MF_{ij} 表示当 A_{ij} 不为零,并且 $i \neq j$ 时 A 的第 i 行第 j 列的元素的值, MF_{ij} 的值为矩阵 A 中第 i 行与第 j 行对应元素不同时为非零或零的元素个数。 MF_{ij} 的值越小表示 i 行(列)与第 j 行(列)的结构越相似。

例如,对矩阵

$$A = \begin{bmatrix} 7 & 0 & 6 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 9 & 3 \end{bmatrix}$$

有 $MF_{13}=4$, $MF_{34}=1$ 。

在选取主元时,选取 MF_{ij} 值最小的元素可以产生较少的填入元。事实上,设旋转矩阵

$$P = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & * & & & & \\ & & & c & s & & \\ & & & * & * & & \\ & & & -s & c & & \\ & & & & & * & \\ & & & & & & 1 \\ & & & & & & & 1 \end{bmatrix}$$

其作用到矩阵 A 上就得到

$$A' = P^T A P,$$

即

$$A' = \begin{bmatrix} & cA_{1i} - sA_{1j} & * & sA_{1i} + cA_{1j} & & & \\ & cA_{2i} - sA_{2j} & * & sA_{2i} + cA_{2j} & & & \\ & * & * & * & & & \\ cA_{1i} - sA_{1j} & * & * & c^2 A_{ii} + s^2 A_{jj} - sc(A_{ij} + A_{ji}) & * & c^2 A_{ij} - s^2 A_{ji} + cs(A_{ii} - A_{jj}) & * & * & cA_{in} - sA_{jn} \\ * & * & * & * & * & * & * & * & * \\ sA_{1i} + cA_{1j} & * & * & c^2 A_{ji} - s^2 A_{ij} + cs(A_{ii} - A_{jj}) & * & c^2 A_{jj} + s^2 A_{ii} + sc(A_{ij} + A_{ji}) & * & * & sA_{in} + cA_{jn} \\ & & & * & * & * & & & \\ & cA_{(n-1)i} - sA_{(n-1)j} & * & sA_{(n-1)i} + cA_{(n-1)j} & & & & & \\ & cA_{ni} - sA_{nj} & * & sA_{ni} + cA_{nj} & & & & & \end{bmatrix}$$

从变换后的矩阵可以看出，当 A 的第 i 行与 A 的第 j 行结构越相似，可能的填入元越少。同理，当 A 的第 i 列与 A 的第 j 列结构越相似，可能的填入元也越少。

3.2 主元选取策略

这里采用的主元选取策略是，先选取比较大的元素集合 **BigSet**，之后在 **BigSet** 中，选取所在行列结构最相似的元素为主元(如果有多个满足条件，则选取这些元素中 A 的值最大者)。在下次选主元时，如果 **BigSet** 不为空，则继续在 **BigSet** 中选主元，直到 **BigSet** 为空或者 (MF_{min} 为 $ROW * COLUMN$)。如果某行已经对角占优，则选较大元素时，该行不予考虑。该过程可以描述如下述算法 1 所示，其中采用文献[8]中的高精度奇异值分解算法，且所涉及的主要变量在表 3.2 中进行了说明。

算法 1. 主元选取

```
While (矩阵A不完全对角占优)
    求RowDominant和较大元矩阵AI, num1为较大元个数;
    While (num1不为零)
        计算AN, MF;
        选AN, AI值为1并且MF值最小的元素为主元。(如果有多个, 则选取A值最大者);
        对A进行相应变化。
        num1=num1-1;
        主元对应的AI变为0;
    If (矩阵A完全对角占优)
        Break;
    EndIf
EndWhile
EndWhile
```

表 1 变量说明表

变量名	变量说明
E	为数值稳定所设的下限值: $E=A(i,i)/(10*ROW*COLUMN)$
N	求较大元时, $N=1, N=2*N$
ROW	为 A 的行数;
Asum	矩阵 A 的元素平方和, 然后再开方
COLUMN	为 A 的列数。
MFmin	MF 中绝对值最小者
AI[ROW][COLUMN]	$AI[i][j]=1$ 表示 A 中的第 i 行第 j 列是较大元, 否则不是。
AN[ROW][COLUMN]	$fabs(A[i][j])>E, AN[i][j]$ 为 1
MF[ROW][COLUMN]	表示第 i 行与第 j 行, 第 i 列与第 j 列结构相异性;
RowDominant[ROW]	A 的第 i 行对角占优时, 为 1, 否则为 0;

4 实验结果与讨论

在实验中，原算法是指 Jin Yun Yuan 和 Plamen Y. Yalamov 提出的算法，policy1 是指在此基础上单纯不考虑已经对角占优的行对应的算法，policy2 对应于算法 1。实验针对 Matrix Market 上的 bcsstk22 等 8 个矩阵对算法进行测试，既有对称稀疏矩阵，也有非对称稀疏矩阵，其结构信息如图 1 所示，概略信息如表 2。

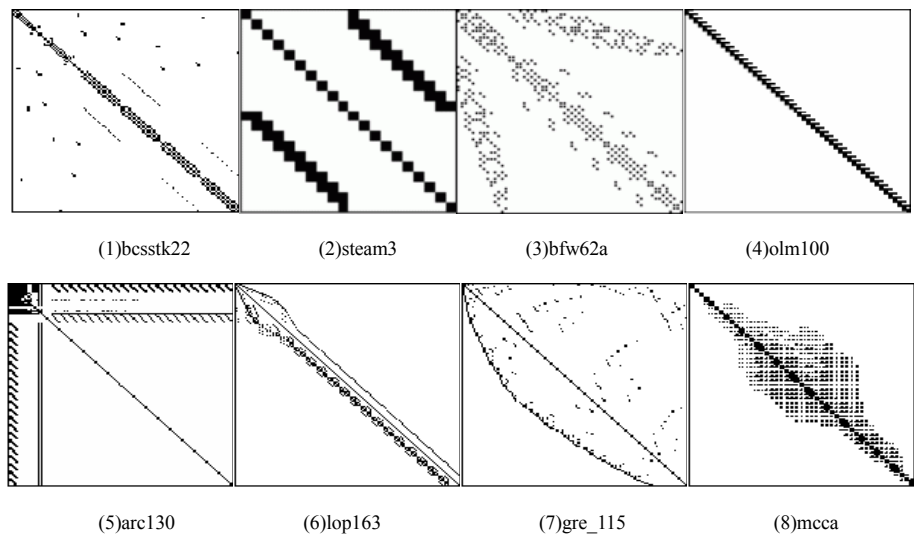


图 1 实验矩阵非零元分布图

表 2 实验矩阵概略信息表

矩阵名	非零元总数	非零对角元	对角线下的非零元个数	对角线上的非零元个数
bcsstk22	696	138	279	279
steam3	314	80	137	97
bfw62a	450	62	191	197
olm100	396	100	148	148
arc130	1037	130	567	340
lop163	935	163	594	178
gre_115	421	115	204	102
mcca	2569	180	1032	1447

在图 2 到图 4 中，分别列出了原算法与文中两种改进算法下的迭代次数、填入元个数、最终所得矩阵中的非零元个数。

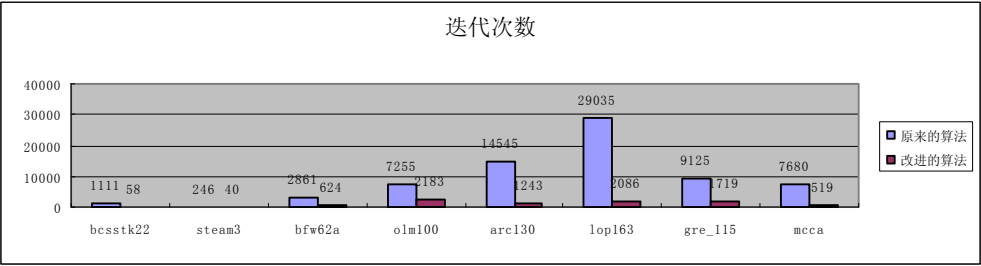


图 2 原来算法与 policy1 迭代次数比较图

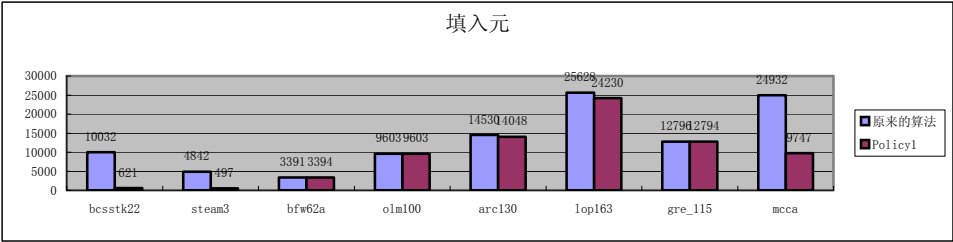


图 3 原来算法与 policy1 填入元比较图

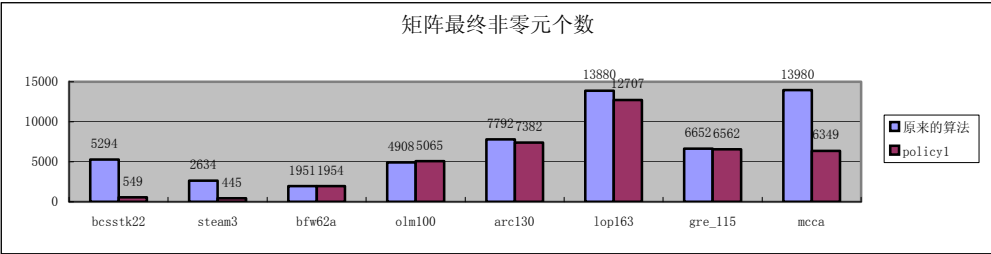


图 4 原来算法与 policy1 最终非零元个数比较图

从图 2 可以看出 policy1 确实能够减少迭代次数。迭代次数的减少，使得选主元的次序发生变化，间接证明不同的选主元次序对迭代次数有影响。从图 3、图 4 可以看出 policy1 在填入元和矩阵最终非零元个数上也有所改进。同时，policy1 相对于原算法，当矩阵是对称矩阵和成块状分布时，改进效果更为明显。之所以如此，一方面是迭代次数减少而减少了填入元个数，从而使得最终所得矩阵中的非零元个数减少，这种减少与矩阵的结构关系很密切。当矩阵不对称或者分布毫无规律而言时，改进效果有可能没那么明显。

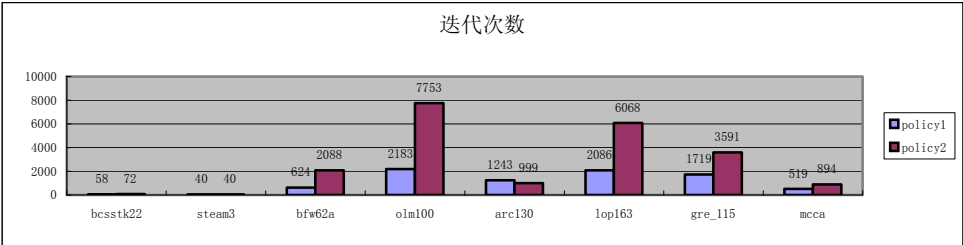


图 5 policy1 与 policy2 迭代次数对比图

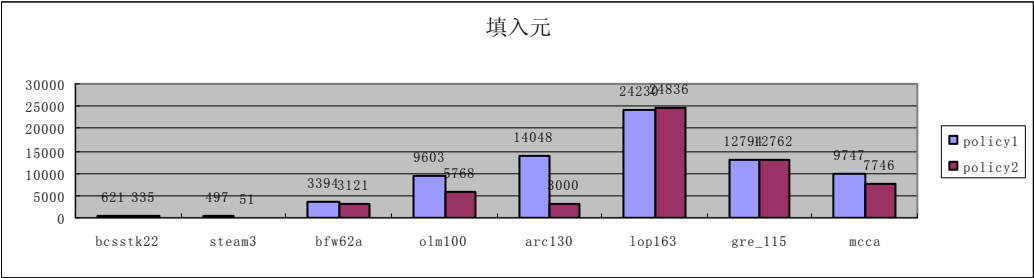


图 6 policy1 与 policy2 过程填入元对比图

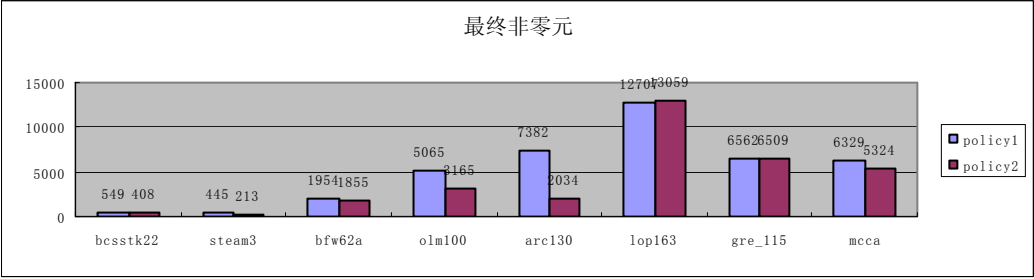


图 7 policy1 与 policy2 最终非零元对比图

policy2 在 policy1 的情况下，考虑了矩阵的行列结构特性，引入了结构相似性，迭代次数虽然比 policy1 有所增加，但是相对于原算法仍然是减少的。而且 policy2 的填入元、最终结果矩阵中的非零元一般都比 policy1 要少。最坏的情况下，如果矩阵结构毫无规律，则该算法在迭代次数上与原算法差不多，但却能够减少最终矩阵的非零元个数。当矩阵成小块状分布，或对称情形下，policy2 的迭代次数比比原算法少得多，并且最终所得矩阵中的非零元个数有所减少。另外 policy2 在迭代过程中产生的填入元一般比 policy1 少。

5 小结

本文提出了一种快速、稳定有效的基于 Jacobi 旋转的稀疏矩阵的对角优势强化方法。该方法能够减少迭代次数、填入元、最终所得矩阵中的非零元个数。算法获得的改进由以下几个方面保证：首先，减少对对角占优性没有影响的迭代；其次，开发稀疏矩阵非零元结构，利用新的主元选取策略降低了填入元的产生。实验结果表明，本文提出的算法相对原算法，迭代次数明显减少，且矩阵稀疏性比原算法保持得更好。

参考文献

[1] 吴建平, 王正华, 李晓梅. 稀疏线性方程组的高效求解与并行计算. 长沙:湖南科学技术出版社, 2007

[2] Olschowka M. and Neumaier A.. A new pivoting strategy for Gauss elimination. Lin. Alg. Appl., 240(1996), 131-151

[3] Duff I. S. and Koster J.. On algorithms for permuting large entries to the diagonal of a sparse matrix. Tech. Report RAL-TR-1999-030. Department for Computation and Information, Atlas Center, Rutherford Appleton

Laboratory, Oxon Ox11 0Qx, 1999

- [4] 吴建平, 王正华, 李晓梅. 块对角占优性与对称矩阵的块对角预条件. 数值计算与计算机应用, 24:4(2003), 241-246
- [5] 吴建平. 稀疏线性代数方程组迭代法中的预处理技术研究. 博士学位论文. 国防科学技术大学计算机学院, 2002
- [6] J.Y. Yuan, Plamen Y. Yalamov. A method for constructing diagonally dominant preconditioners based on Jacobi rotation. Appl. Math. Comp., 174(2006), 74-80
- [7] 杨绍祺, 谈跟林. 稀疏矩阵-算法及程序实现. 北京:高等教育出版社, 1985
- [8] 徐士良. C 常用算法程序集. 北京:清华大学出版社, 2004

作者简介

银福康 (1984 -), 男, 广西象州, 硕士生,主要研究科学计算与大气遥感数据处理.

宋君强 (1962 -), 男, 湖南宁乡, 博导, 研究员, 主要从事高性能计算和数值天气预报.

吴建平 (1974 -), 男, 湖南新化, 博士, 副研究员, 主要从事科学计算与并行算法方面的研究。

服从二维指数分布的非独立随机变量的 线性组合的分布

郭云飞¹ 尹哲^{1,2}

(1. 延边大学数学系, 延吉 133002; 2. 北京大学信息管理系, 北京 100871)

摘 要: 国内外学者对 $\alpha X + \beta Y$ 的分布的研究很多, 然而大部分都是在 X 与 Y 独立并且服从同一分布的前提下研究的, 而对 X 与 Y 非独立的情况研究很少, 至今未在国内见到相关研究成果, 本文将基于这种考虑, 以在可靠性中应用最广泛分布之一的二维指数分布为例, 推出了 $\alpha X + \beta Y$ 的分布。本文是受可靠性及质量工程等方面的现实例子启发下完成的。

关键词: 二维指数分布; 非独立; 线性组合

Exact Distributions for the linear combination of Bivariate Exponential Components of Dependent Variable

GUO Yun-fei¹ YIN Zhe^{1,2}

(1. Mathematics Department, Yanbian University, Yanji 133002, China;

(2. Department of Information Management, Keking University, Beijing 100871, China)

Abstract: The distribution of $\alpha X + \beta Y$ has been studied by several authors especially when X and Y are independent random variables and come from the same family. However, there is relatively little work of this kind when X and Y are correlated random variables. I haven't found related results in domestic so far. Based on this consideration, in this paper, we take bivariate exponential distribution which is widely applied in reliability as an example, derive the exact distributions of $\alpha X + \beta Y$. The work is motivated by real-life examples in quality and reliability engineering.

Keywords: Bivariate exponential distributions; dependent; linear combination

1 引言

对于给定的随机变量 X 和 Y , $\alpha X + \beta Y$ 这种线性组合形式的分布在质量工程和可靠性工程中有重大意义^[1], 例如: 新鲜水果和蔬菜的质量决定在农业生产中非常重要。尽量发展了大量的技术去对水果和蔬菜的质量进行非破坏性的评估, 质量分类的方法主要是靠手工决

定。*Ozer*^[2]改进了模型，把水果的质量定义成大量参数的线性组合，根据这些参数可以给水果进行分类。当处理两个或更多的控制变量时，人们总会对这些变量最佳线性组合感兴趣，在这方面也有很多结果，例如 *Glynn*^[3]。已经有很多学者，尤其是在 X 与 Y 独立的前提下研究了 $\alpha X + \beta Y$ 的分布，例如，*Fisher*^[4] 和 *Chapman*^[5] 研究了 t 分布，*Christopeit and Helmes*^[6] 研究了正态分布，*Davies*^[7] 和 *Farebrother*^[8] 研究了 χ^2 分布，*Ali and Obaidullah*^[9] 研究了指数分布，*Moschopoulos*^[10] 和 *Provost*^[11] 研究了 Γ 分布，*Dobson*^[12] 研究了泊松分布。然而，当 X 与 Y 非独立时对 $\alpha X + \beta Y$ 分布的研究几乎没有成果，国内至今未见到，国外见到的也是屈指可数的几篇，而本文将会考虑当 X 与 Y 服从联合生存概率^[13]

$$\overline{F}(x,y)=\exp\{-\lambda_1x-\lambda_2y-\lambda_{12}\max(x,y)\},\lambda_1,\lambda_2,\lambda_{12},x,y>0 \tag{1.1}$$

此分布被称为 $MOBVE(\lambda_1,\lambda_2,\lambda_{12})$ ，还可以写出他的联合概率密度函数^[14]

$$f(x,y)=\begin{cases} \lambda_1(\lambda_2+\lambda_{12})\exp\{-\lambda_1x-(\lambda_2+\lambda_{12})y\},x<y \\ \lambda_2(\lambda_1+\lambda_{12})\exp\{-\lambda_2y-(\lambda_1+\lambda_{12})x\},x>y \\ \lambda_{12}\exp\{-(\lambda_1+\lambda_2+\lambda_{12})y\},x=y \end{cases} \tag{1.2}$$

此分布有着广泛的应用，特别是在可靠性理论中。

2 PDFS

在 $Z=\alpha X+\beta Y$ 中，常数 α ， β 可正可负，所以有四种可能： $\alpha>0,\beta>0;\alpha>0,\beta<0;\alpha<0,\beta>0;\alpha<0,\beta<0$ ，但从对称的角度考虑，我们有充分理由只考虑两种情况： $\alpha>0,\beta>0;\alpha<0,\beta>0$ 。由于 $x=y$ 时 $f(x,y)$ 转化为一元函数，因此我们在这里只考虑 $x\neq y$ ， $x<y$ 或 $x>y$ 的情况。下面我们将推导这两种情况下 $Z=\alpha X+\beta Y$ 的概率密度函数 PDFS。

定理 1 如果 X 和 Y 服从联合概率密度函数 (1.2)，那么当 $\alpha>0,\beta>0$ 时 $Z=\alpha X+\beta Y$ 的 PDFS 为：

$$f_s(s)=\begin{cases} \frac{\lambda_1(\lambda_2+\lambda_{12})}{-\lambda_1\beta+(\lambda_2+\lambda_{12})\alpha}\left(\exp\left\{-\frac{\lambda_1}{\alpha}s\right\}-\exp\left\{-\frac{\lambda_2+\lambda_{12}}{\beta}s\right\}\right),x<y \\ \frac{\lambda_2(\lambda_1+\lambda_{12})}{\lambda_2\alpha-(\lambda_1+\lambda_{12})\beta}\left(\exp\left\{-\frac{\lambda_1+\lambda_{12}}{\alpha}s\right\}-\exp\left\{-\frac{\lambda_2}{\beta}s\right\}\right),x>y \end{cases} \quad 0<s<\infty \tag{2.1}$$

证明：根据 (1.2)，我们有：

(a) $x<y$ 时，

$$f_s(s)=\frac{1}{\beta}\int_0^{\frac{s}{\alpha}}f\left(x,\frac{s-\alpha x}{\beta}\right)dx$$

$$\begin{aligned}
&= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \int_0^{\frac{s}{\alpha}} \exp\left\{-\lambda_1 x - (\lambda_2 + \lambda_{12}) \frac{(s - \alpha x)}{\beta}\right\} dx \\
&= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\} \int_0^{\frac{s}{\alpha}} \exp\left\{\frac{[-\lambda_1 \beta + (\lambda_2 + \lambda_{12}) \alpha] x}{\beta}\right\} dx \\
&= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \frac{\beta}{-\lambda_1 \beta + (\lambda_2 + \lambda_{12}) \alpha} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\} \int_0^{\frac{s}{\alpha}} d \exp\left\{\frac{[-\lambda_1 \beta + (\lambda_2 + \lambda_{12}) \alpha] x}{\beta}\right\} \\
&= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{-\lambda_1 \beta + (\lambda_2 + \lambda_{12}) \alpha} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\} \left[\exp\left\{\left(-\frac{\lambda_1}{\alpha} + \frac{\lambda_2 + \lambda_{12}}{\beta}\right) s\right\} - 1\right] \\
&= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{-\lambda_1 \beta + (\lambda_2 + \lambda_{12}) \alpha} \left(\exp\left\{-\frac{\lambda_1}{\alpha} s\right\} - \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\}\right)
\end{aligned}$$

(b) $x > y$ 时

$$\begin{aligned}
f_S(s) &= \frac{1}{\beta} \int_0^{\frac{s}{\alpha}} f\left(x, \frac{s - \alpha x}{\beta}\right) dx \\
&= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \int_0^{\frac{s}{\alpha}} \exp\left\{-(\lambda_1 + \lambda_{12}) x - \lambda_2 \frac{(s - \alpha x)}{\beta}\right\} dx \\
&= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\} \int_0^{\frac{s}{\alpha}} \exp\left\{\frac{[\lambda_2 \alpha - (\lambda_1 + \lambda_{12}) \beta] x}{\beta}\right\} dx \\
&= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \frac{\beta}{\lambda_2 \alpha - (\lambda_1 + \lambda_{12}) \beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\} \int_0^{\frac{s}{\alpha}} d \exp\left\{\frac{[\lambda_2 \alpha - (\lambda_1 + \lambda_{12}) \beta] x}{\beta}\right\} \\
&= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\lambda_2 \alpha - (\lambda_1 + \lambda_{12}) \beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\} \left[\exp\left\{\left(-\frac{\lambda_1 + \lambda_{12}}{\alpha} + \frac{\lambda_2}{\beta}\right) s\right\} - 1\right] \\
&= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\lambda_2 \alpha - (\lambda_1 + \lambda_{12}) \beta} \left(\exp\left\{-\frac{\lambda_1 + \lambda_{12}}{\alpha} s\right\} - \exp\left\{-\frac{\lambda_2}{\beta} s\right\}\right)
\end{aligned}$$

即证明了 (2.1)。

定理 2 如果 X 和 Y 服从联合概率密度函数 (1.2), 那么当 $\alpha < 0, \beta > 0$ 时 $Z = \alpha X + \beta Y$ 的 PDF 为:

$$f_S(s) = \begin{cases} \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\lambda_1 \beta - (\lambda_2 + \lambda_{12}) \alpha} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\}, & x < y \\ \frac{\lambda_2(\lambda_1 + \lambda_{12})}{-\lambda_2 \alpha + (\lambda_1 + \lambda_{12}) \beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\}, & x > y \end{cases} \quad 0 < s < \infty \quad (2.2)$$

$$f_S(s) = \begin{cases} \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\lambda_1 \beta - (\lambda_2 + \lambda_{12}) \alpha} \exp\left\{-\frac{\lambda_1}{\alpha} s\right\}, & x < y \\ \frac{\lambda_2(\lambda_1 + \lambda_{12})}{-\lambda_2 \alpha + (\lambda_1 + \lambda_{12}) \beta} \exp\left\{-\frac{\lambda_1 + \lambda_{12}}{\alpha} s\right\}, & x > y \end{cases} \quad s < 0 \quad (2.3)$$

证明: 若 $s > 0$, 且 $x < y$ 时, 有:

$$\begin{aligned}
 f_s(s) &= \frac{1}{\beta} \int_0^\infty f\left(x, \frac{s-\alpha x}{\beta}\right) dx \\
 &= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \int_0^\infty \exp\left\{-\lambda_1 x - (\lambda_2 + \lambda_{12}) \frac{(s-\alpha x)}{\beta}\right\} dx \\
 &= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \frac{\beta}{-\lambda_1 \beta + (\lambda_2 + \lambda_{12})\alpha} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\} \int_0^\infty d \exp\left\{\frac{[-\lambda_1 \beta + (\lambda_2 + \lambda_{12})\alpha]x}{\beta}\right\} \\
 &= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\lambda_1 \beta - (\lambda_2 + \lambda_{12})\alpha} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\}
 \end{aligned}$$

若 $s > 0$, 且 $x > y$ 时, 有:

$$\begin{aligned}
 f_s(s) &= \frac{1}{\beta} \int_0^\infty f\left(x, \frac{s-\alpha x}{\beta}\right) dx \\
 &= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \int_0^\infty \exp\left\{-(\lambda_1 + \lambda_{12})x - \lambda_2 \frac{(s-\alpha x)}{\beta}\right\} dx \\
 &= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \frac{\beta}{\lambda_2 \alpha - (\lambda_1 + \lambda_{12})\beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\} \int_0^\infty d \exp\left\{\frac{[\lambda_2 \alpha - (\lambda_1 + \lambda_{12})\beta]x}{\beta}\right\} \\
 &= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{-\lambda_2 \alpha + (\lambda_1 + \lambda_{12})\beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\}
 \end{aligned}$$

若 $s < 0$, 且 $x < y$ 时, 有:

$$\begin{aligned}
 f_s(s) &= \frac{1}{\beta} \int_{\frac{s}{\alpha}}^\infty f\left(x, \frac{s-\alpha x}{\beta}\right) dx \\
 &= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \int_{\frac{s}{\alpha}}^\infty \exp\left\{-\lambda_1 x - (\lambda_2 + \lambda_{12}) \frac{(s-\alpha x)}{\beta}\right\} dx \\
 &= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\beta} \frac{\beta}{-\lambda_1 \beta + (\lambda_2 + \lambda_{12})\alpha} \exp\left\{-\frac{\lambda_2 + \lambda_{12}}{\beta} s\right\} \int_{\frac{s}{\alpha}}^\infty d \exp\left\{\frac{[-\lambda_1 \beta + (\lambda_2 + \lambda_{12})\alpha]x}{\beta}\right\} \\
 &= \frac{\lambda_1(\lambda_2 + \lambda_{12})}{\lambda_1 \beta - (\lambda_2 + \lambda_{12})\alpha} \exp\left\{-\frac{\lambda_1}{\alpha} s\right\}
 \end{aligned}$$

若 $s < 0$, 且 $x > y$ 时, 有:

$$\begin{aligned}
 f_s(s) &= \frac{1}{\beta} \int_{\frac{s}{\alpha}}^\infty f\left(x, \frac{s-\alpha x}{\beta}\right) dx \\
 &= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \int_{\frac{s}{\alpha}}^\infty \exp\left\{-(\lambda_1 + \lambda_{12})x - \lambda_2 \frac{(s-\alpha x)}{\beta}\right\} dx \\
 &= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{\beta} \frac{\beta}{\lambda_2 \alpha - (\lambda_1 + \lambda_{12})\beta} \exp\left\{-\frac{\lambda_2}{\beta} s\right\} \int_{\frac{s}{\alpha}}^\infty d \exp\left\{\frac{[\lambda_2 \alpha - (\lambda_1 + \lambda_{12})\beta]x}{\beta}\right\} \\
 &= \frac{\lambda_2(\lambda_1 + \lambda_{12})}{-\lambda_2 \alpha + (\lambda_1 + \lambda_{12})\beta} \exp\left\{-\frac{(\lambda_1 + \lambda_{12})}{\alpha} s\right\}
 \end{aligned}$$

这就完成了定理2的证明。

3 结论

我们推导了当 X 和 Y 服从联合概率密度函数(1.2)时各种情况下 $Z = \alpha X + \beta Y$ 的分布,这对于实践中有着重要的意义,特别是在可靠性中,我们可以考虑多个部件并联或串联的寿命分布,当然我们会在今后的研究中讨论 X 和 Y 服从其他类型的分布时 $Z = \alpha X + \beta Y$ 的分布情况。

参考文献

- [1] Arjun K.Gupta and Saralees Nadarajah . Exact and approximate distributions for the linear combination of inverted Dirichlet components.[J] J.Japan Staist.Soc ,2006 36(2):225-236.
- [2] Ozer, N., Engel, B. A. and Simon, J. E. A multiple impact approach for non-destructive measurement of fruit firmness and maturity, Transactions of the ASAE, 1998,41, 871-876.
- [3] Glynn, P. W. and Iglehart, D. L. The optimal linear combination of control variates in the presence of asymptotically negligible bias, Naval Research Logistics Quarterly, 1989,36,683-692.
- [4] Fisher, R. A. The fiducial argument in statistical inference, Annals of Eugenics, 1935,6,391-398.
- [5] Chapman, D. G. Some two sample tests, Annals of Mathematical Statistics, 1950,21, 601-606.
- [6] Christopheit, N. and Helmes, K. A convergence theorem for random linear combinations of independent normal random variables, Annals of Statistics, 1979,7, 795-800.
- [7] Davies, R. B. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables, Applied Statistics, 1980,29, 323-333.
- [8] Farebrother, R. W. Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables, Applied Statistics, 1984,33, 332-339.
- [9] Ali, M. M. and Obaidullah, M. Distribution of linear combination of exponential variates, Communications in Statistics—Theory and Methods, 1982,11, 1453-1463.
- [10] Moschopoulos, P. G. The distribution of the sum of independent gamma random variables, Annals of the Institute of Statistical Mathematics, 1985,37, 541-544.
- [11] Provost, S. B. On sums of independent gamma random variables, Statistics, 1989,20, 583-591.
- [12] Dobson, A. J., Kulasmaa, K. and Scherer, J. Confidence intervals for weighted sums of Poisson parameters, Statistics in Medicine, 1991,10, 457-462.
- [13] 程侃, 曹晋华 可靠性数学引论[M], 科学出版社,1986.
- [14] Saralees Narajah and Samuel Kotz. Reliability for some Bivariate exponential distributions, Mathematical Problems in Engineering,2006,1-14.

作者简介

郭云飞 (1983—), 男, 延边大学数学系硕士研究生, 从事统计方向研究。
Email:guoyunfei0413@sina.com

尹哲 (1963—), 延边大学数学系, 从事统计计算方向的研究。Email: yinzhe@ybu.edu.cn

一种PDF信息提取与表格重现的算法

张 伯¹ 陈 彩²

(1. 北京工业大学 计算机学院,北京市 100124;

2. 北京工业大学 计算机学院,北京市 100124)

摘 要: PDF 是一种国际通用格式的电子文档,与传统扫描图像和流行于网络的标记语言文档相比,PDF 表格既无明确的实体框架信息,也没有采用结构化语言进行描述,这给 PDF 文件中表格信息的提取、复用和编辑带来了诸多不便。本文提出并实现了一种使 PDF 文档表格的逻辑结构得以重现的算法,并将表格内容以 HTML 序列化输出。该算法为 PDF 表格信息的再利用提供了便利。

关键字: 信息提取; 表格重现; PDF

An Algorithm for Information Extraction and Recognition of Tables from PDF

Abstract: The PDF is the international general electronic document format. Neither information of explicit entity framework nor structural description language exists in PDF format to define the tables, which brought many inconvenience to extracting, duplicating and editing of the information of tables in PDF. This article provides an algorithm to represent the logic structure of tables of PDF document, which can output the content of tables in HTML format. This algorithm brought convenience to the reuse of table information of PDF.

Keywords: Information Extraction; Table Representation; PDF

1 引言

PDF (Portable document format) 是一种目前国际通用的电子文档开放标准^[1],其平台无关、信息完整、安全可靠等特性备受关注,各国政府机关、企事业单位、出版行业均大量采用该格式作为标准,进行信息发布、交换与存储。与此同时,对 PDF 文档信息的提取、复用和再编辑的需求也愈发强烈,表格便是其中重要内容之一。PDF 表格重现与传统电子表格相比,识别过程具有一定特殊性。

从理论上讲,表格是在人的视觉经验下看似横平竖直排列的一些文字。因此表格识别也是基于人的视觉经验的。从编码格式看,PDF 并没有专门提供表格信息记录的编码方式,表格线是以底纹图形方式给出的,无法跟文字进行明确的逻辑关联,也就是说,只有在 PDF 可视化后,表格存在与否才能明确,所以表格线只具有参考意义。

OCR, 即光学字符识别技术^[2]出现较早, 它利用扫描图像各点灰度不同, 来判断文字、边框、照片等信息。对于边框结构信息完整的表格, 扫描图像技术可以通过识别表格线的交叉点以及灰度的均匀程度来判断表格线以及其与文字的关系^[2,3]。而对于边框结构不完整的表格, 也有利用灰度频率、字块密集程度等信息来识别表格内容及其逻辑布局关系的^[4,5]。由于扫描图像与 PDF 特征的本质不同, 识别算法很难借鉴到 PDF 内容的识别技术中。

HTML——超文本标记语言^[6]目前被广泛应用于网络, 其编码格式具有天然的结构化特点, 只要找到表格便签便可以侦测到表格位置大小等各种信息。相比之下, 绝大部分 PDF 中的表格线是以图像形式给出, 也有很少一部分采用矢量线描述, 线框信息与文字很难结合起来, 使得信息抽取过程存在诸多问题。

有文献曾提出对 PDF 中表格的提取可以先转化为图像^[7], 这样很好地利用了现有研究成果。不过这种方式仍值得商榷^[8], 因为间接识别出表格是以损失丰富的原始数据信息为代价^[9,10]。

本文提出了一种针对 PDF 信息提取和表格重现算法, 该算法也适用于其他具有无结构化编码特征的表格的重现。

2 处理流程概要

处理流程如图 1 所示。

- Step 1: PDF 文档解析

根据 PDF 编码规则, 对二进制码流进行解码, 从内容流中分离出文本、图像等信息。

- Step 2: 文字流生成与框选内容抽取

建立文字流数据结构并保存文本对象信息, 将 PDF 内容可视化, 框选待重现表格内容。

- Step 3: 栅格化

对所有文字流节点分别按照水平和垂直方向进行划分, 将划分信息分别保存, 形成概念上的待重现表格边框结构。

- Step 4: 表格内容归位

将仅有坐标而无结构化信息的文字流内容, 根据概念结构找到自己在表格中相对位置, 从而建立了文字流节点间的相对关系, 实现表格的拓扑结构。

- Step 5: 序列化输出

最后对建立好拓扑结构的二维表格进行一维序列化输出, 表示为通用结构化编码格式, 如 HTML 等, 可以在网页中进行浏览, 或导出到 OA 软件中进行可视化编辑。

3 PDF文档解析

PDF 文档通过文档描述语言进行编码。在该语言中, 文档由二维对象集合组成, 比如文本、图元、图像等, 这些对象包含在内容流中。每个对象都有一些元数据, 包含了表现特征和位置坐标信息, 保证了页面的准确显示和打印输出。对象在内容流中不规则的排列, PDF 内容只能在绘制或者打印后才能被理解。

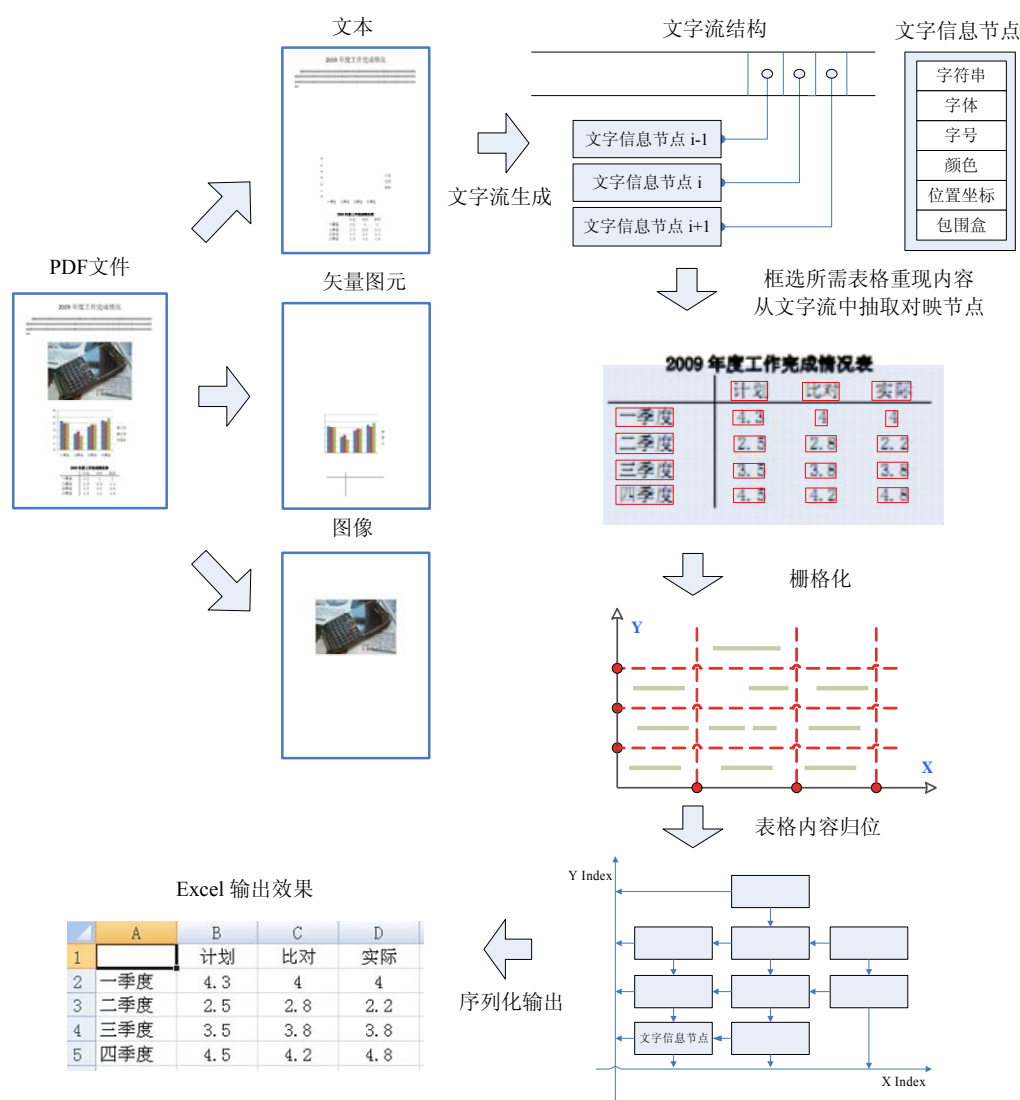


图 1 整体处理流程

规范的 PDF 文档物理结构由四部分组成，如图 2 所示。

- 文件头 (Head)：指示该文件所属 PDF 规范的版本号，出现在 PDF 文件第一行
- 文件体 (Body)：包含了一系列对象，组成了 PDF 文件的主要部分
- 交叉引用表 (Xref Table)：包含了文件中的间接对象信息
- 文件尾 (Trailer)：给出了交叉引用表，并且可以明确获取各对象位置

PDF 是一种逻辑清晰严密地结构化文件格式，如图 3 所示，由诸多被称之为“对象 (Object)”的模块组成。每个对象都有不同且唯一的数字标号，这样使得其他对象可以方便引用。在 PDF 文档中，对象出现的顺序是没有语义的。一般一个应用可以通过引用从一个对象关联到另一个对象，而非对于所有对象顺序处理。这对于文档交互和随机访问非常重要。

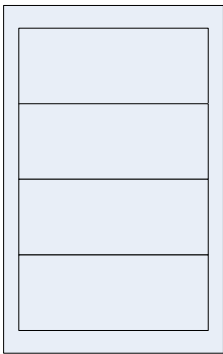


图 2 PDF 文档物理结构

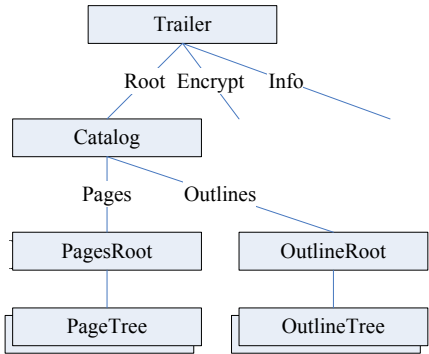


图 3 PDF 文档逻辑结构

想要将所有 PDF 对象关联起来，首先要找到根对象（Root），而根部的信息在文件尾（Trailer）中进行了描述，同时被描述的还有文档信息字典和密码字典。找到了根对象也就明确了交叉引用表的位置，通过查询交叉引用表还可以找到目录对象（Catalog）。目录对象包含文档大纲（Outline）和页面组对象（Pages）。而前者列出了 PDF 的书签树，后者则包含了所有页面对象的引用。

4 文字流生成与框选内容抽取

PDF 文档中的每页中包含了一个或多个内容流。内容流打包成一个包含指令序列的绘制元素，如 X 对象，底纹，字体和注释，其中 X 对象又包括路径对象、文本对象和样本图像。为了便于表格识别，本算法根据 PDF 语法提取出文字相关的信息，建立文字流结构，即文字信息节点链，如图 4、5 所示，其中每个节点包含一个 PDF 原子字符串以及相关信息，如位置坐标，包围盒信息等。对于非文本部分，同样建立类似的流结构，以便屏幕显示、打印输出等操作。

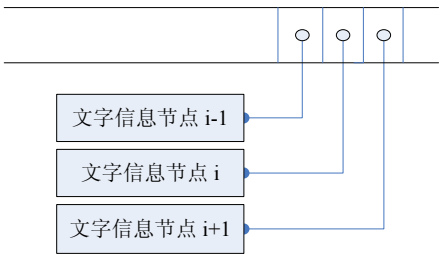


图 4 文字流结构



图 5 文字信息节点

PDF 编码格式中提供文本状态参量和操作符，如 Tf 表示文本字体、Tw 表示字符间距等等，还可以从原始数据中提取到页面起始地址。为了便于表格识别，需要更为精确的文本特征信息——原子字符串包围盒和单字符包围盒信息，而这些信息无法从编码中直接获取，需要综合各种文本状态参量计算得出。

PDF 提供了比较丰富的文本状态参量信息，文本对象定位信息主要分为三类：第一类是字体（Font）特征信息，决定了文本的内在特征，第二类是字符特征信息，比如字号，放缩比

例等，表示了文本内单个字符所占空间信息；第三类是字间特征信息，比如字符间距，相对位置等，表示了文本内所有字符的布局关系。综合这三类信息，可以精确计算出文本整体以及单个字符的包围盒数据。

在不同的平台上，以 Window 为例，可以通过 GDI 或 Cairo 等 2D 图形绘制引擎将 PDF 内容进行呈现，同时将设置适当的交互功能方便使用者框选所需抽取的表格内容，如图 6 所示。由于文字流中字符串是以单元形式整体出现，而不同行的字串起始和终止位置又不尽相同，这样可能导致框选区域边界与一部分独立的字符串相交。考虑到字符串一般表示一个相对完整的语义，本算法将剔除这些压线字符串，只保留处于框选区域之中的完整字符串。从实际效果看，这样操作更加符合人的认知。

2009 年度工作完成情况表			
	计划	对比	实际
一季度	4.3	4	4
二季度	2.5	2.8	2.2
三季度	3.5	3.8	3.8
四季度	4.5	4.2	4.8

图 6 框选所需抽取内容

5 栅格化

本算法将采用多趟处理的方式对表格进行重现。第一步是栅格化，主要任务是确定划线位置。由于带有边框视觉效果的底纹图像无法被利用为有效地边框信息来关联处理表格识别，待重现的框选字符串实际成了一个个离散的文字团，每个文字团占据了一个概念中的表格单元。分别从水平方向和垂直方向上看，文字团之间又存在着同行或共列现象。要将这种概念形式转化为真正表格，首先要做的是确定表格线的位置。

具体做法如下：取当前节点包围盒边界，以 X 方向的左右边界为例，当前节点左右边界投影是否击中其他文字区域，如图 7 所示，如果两节点邻接边界距离过近，视为被击中。遍历后，将从未击中过其他节点区域的边界按照升序方式一一插入 X 方向分段信息链，实现对 X 空间维度上的切割，Y 空间维度亦然，如图 8 所示。

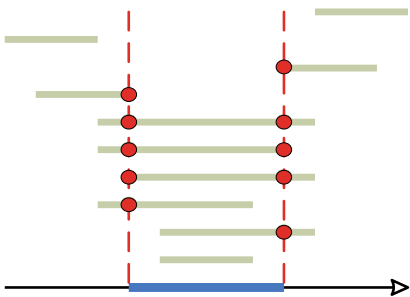


图 7 击中图

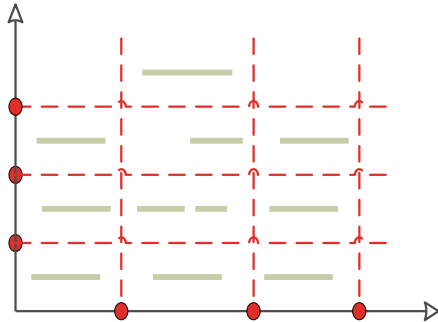


图 8 栅格化示意图

6 表格内容归位

第一趟生成了栅格化信息，初步确定了表格框架位置和文字流节点间的逻辑位置关系，然而此成果还停留在概念表格阶段。为了方便的操作文字流节点，最终实现 HTML 的序列化输出，第二趟处理将结合 HTML 语法特征，对文字流节点信息进行二维重构，该过程也是表格内容归位的过程，将生成完整的二维表结构，如图 9 所示。

HTML 语法中，Tr、Td 两种标签分别描述表格的行和列信息。鉴于此，本算法将采用二维表形式描述表格结构。首先建立表对象。表对象中包含了一个行对象的有序序列，每个行对象中又包含了列对象有序序列。在步骤四中，通过栅格化操作已经明确了行列数量，随后将栅格化所形成的概念表格单元一一映射到二维表格框架中，并将文字流中各节点内容分别归位到框架单元中，对概念表格进行实体化。此时各单元格的内容、空间位置以及单元格间相对位置关系已然明晰确定。

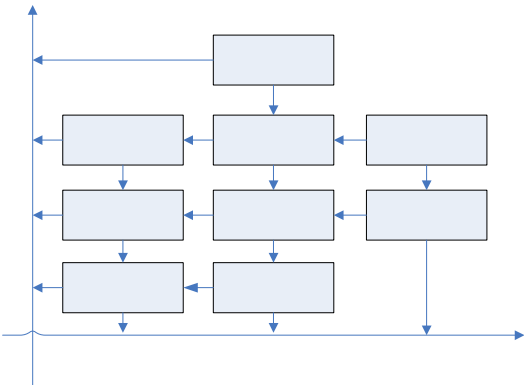


图 9 二维表格框架

7 序列化输出

经过栅格化、二维框架生成以及内容归位操作，一个具有完整结构化信息的表格已然形成。然而目前的结构通用性不强，无法在 OA 软件中读取。为了使表格内容方便复用与编辑，需将二维表结构导出为更为通用的标准格式，如 HTML、XML、CVS 等。以 HTML 为例，首先取第一行，从左到右依次遍历列节点，拼装出 TD 标签单元，将行中所有单元依次连接，首尾拼装 TR 标签单元，之后进行第二行处理，以此类推直至最后一行结束。

8 结论与未来工作

本文实现了一套针对 PDF 表格，即无结构化信息表格的信息抽取与重现算法。首先，从 PDF 内容流中分离生成文字流。其次，为了提高表格内容选取的准确性并丰富文字流内容，

计算出流节点中字符串和单个字符的包围盒信息。再次，从文字流中抽取框选确定的表格区域所对映的文字流节点，对抽取结果按照水平和垂直两个方向分别进行栅格化处理，将文字流节点形成的文字团划分出概念表格。再次，根据栅格化后两个方向的分割线形成以二维表为存储结构的实体表格，文字流节点内容将对号入座到表格单元中。由此原本无结构化信息的表格便具有明确的结构化特征。最后，系统将这一结果导出为 HTML 格式以便复用和编辑。

由于合并格存在，表格格式产生了很大不确定性^[11]，从大量的 PDF 文档中也可可见一斑，本算法目前仅用于不含合并格的表格的处理，合并格处理也是本算法未来需要拓展的重要内容。对于大量出现且结构相同的复杂表格，可以针对其特征定制处理，这样保证重现较高的精确性和可用性。想要得到复杂表格重现的精确结果，结合人工操作是个不错的选择。

参考文献

- [1] <http://www.pdf.net.cn/>
- [2] Mori, C.Y. Suen and K. Yamamoto, Historical Review of OCR Research and Development, Proc. IEEE, vol. 80, no. 7, pp. 1,029-1,058 (1992).
- [3] S. Mandal, S. P. Chowdhury, A. K. Das (2006) A simple and effective table detection system from document images. In: International Journal of Document Analysis 8(2): 172–182 (2006)
- [4] Ramel, J.-Y., Crucianu, M., Vincent, N., Faure, C.: Detection, extraction and representation of tables. In: 7th International Conference on Document Analysis and Recognition, vol. 1, pp. 374–378. Edinburgh, UK (2003)
- [5] Kieninger, T.: Table structure recognition based on robust block segmentation. In: V Document Recognition, Proceedings of SPIE. San Jose, USA (1998)
- [6] HTML 4.01 Specification W3C Recommendation 24 December 1999
- [7] Embley, D.W., Lopresti, D., Nagy, G.: Notes on Contemporary Table Recognition. In: Proc. 7th International Workshop on Document Analysis Systems (DAS 2006), Nelson, New Zealand, pp. 164–175 (2006)
- [8] H. Wasserman, K. Yukawa, B. Sy, K. Kwok, I.T. Phillips, A theoretical foundation and a method for document table structure extraction and decomposition, in: D. Lopresti, J. Hu, R. Kashi (Eds.), Document Analysis Systems V, Fifth IAPR International Workshop on Document Analysis Systems, Princeton, NJ, USA, August 2002, pp. 291–294.
- [9] H. Chao and J. Fan. Layout and Content Extraction from PDF Documents. In Proc. of DAS 2004, 6th IAPR Workshop on Document Analysis Systems, pages 213--224, Florence, Italy, September 2006.
- [10] Yildiz, B., Kaiser, K., Miksch, S.: pdf2table: A method to extract table information from PDF files. In: 2nd Indian Int. Conf. on AI, Pune (2005)
- [11] Xinxin Wang, Derick Wood, Tabular abstraction, editing, and formatting (1996)

作者简介

张伯，男，1982 年，硕士研究生，研究领域：计算机软件与理论；

陈彩，女，1963 年，副教授，研究领域：计算机软件与理论。

An Estimate Method for the Maximum Maintenance Time of Normal Distribution Items

JIN Xing PENG Bo LU Hai

(The Academy of Equipment Command & Technology, P.O.Box: Beijing 3380-86, 101416, China)

Abstract: The maintenance plan can be prepared according to the maximum maintenance time, that is calculated by test data of maintenance sampling. Based on the non-central t -distribution, the equation of upper-bound and lower-bound of maximum maintenance time is deduced. An estimate method of confidence interval of maximum maintenance time is presented. Furthermore an unbiased estimate method of maximum maintenance time is proposed, which can be applied any sample size. The theory of design and analysis can be provided by the method.

Keywords: Normal Distribution; Sampling Test; Maintenance Time

In the engineering, it is urgent to figure out the method for determining the maximum maintenance time of the normal distribution items based on the limited sample maintenance test data (namely the upper-bound value and the lower-bound value), get the confidence interval and the unbiased estimate of the maximum maintenance time consequently.

1 Maximum Maintenance Time of Normal Distribution Items

Let T be the random variable representing the item maintenance time and normally distributed $T \sim N(\mu, \sigma^2)$. The maintainability is as follows:

$$M(t) = P(T \leq t) = \Phi\left(\frac{t - \mu}{\sigma}\right) \quad (1)$$

Where the mean and standard deviation of maintenance time is μ and σ , $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

Let maintainability $M(t_p) = p$, when the probability p is specified, the corresponding maintenance time t_p is

$$t_p = \mu + u_p \sigma \quad (2)$$

Where u_p is the lower quantile with specified probability p of the standard normal distribution. If we let the maintainability $M(t_{0.9}) = 0.9$ where $p = 0.9$, $t_{0.9}$ is the maximum maintenance time. Certainly we can let $p = 0.95$ or $p = 0.99$, so $t_{0.95}$ or $t_{0.99}$ is the maximum maintenance time.

2 Upper-bound and Lower-bound of Maximum Maintenance Time

The mean and the standard deviation of the maintenance time samples are as follows:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n t_i \quad S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (t_i - \bar{X})^2} \tag{3}$$

Where $t_i (i=1,2,L,n)$ is the sample value, n is the number of sample.

If the maintenance probability p is given, the maintenance time $t_p = \mu + u_p \sigma$ is just theoretical. So based on the value of limited item maintenance time sample $t_i (i=1,2,L,n)$, we can only get the upper-bound t_1 and the lower-bound t_2 expressed in the random variable form below.

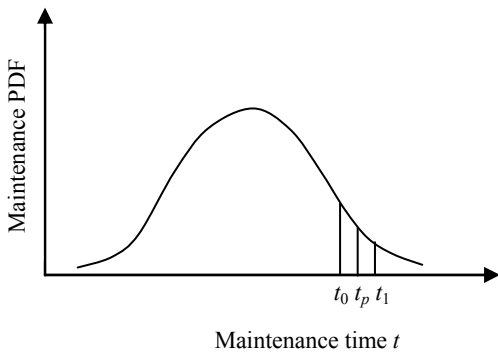


Figure1 sketch of the maintenance time under given probability value

$$t_1 = \bar{X} + K_1 S \quad t_0 = \bar{X} + K_0 S \tag{4}$$

Where K_0 and K_1 are constant value and $K_1 > K_0$. See Fig. 1.

3 Method for Defining Upper & Lower-bound of Maximum Maintenance Time

To get the better estimate of the maximum maintenance time $t_p = \mu + u_p \sigma$, let the upper-bound of the maximum meet

$$P(t_1 < t_p) = \alpha / 2 \tag{5}$$

And let the lower-bound of the maximum $t_0 = \bar{X} + K_0 S$ meet

$$P(t_0 > t_p) = \alpha / 2 \tag{6}$$

To guarantee the maximum maintenance time t_p be in the interval $[t_0, t_1]$ with the given probability $1 - \alpha$ more likely, then

$$P(t_0 < t_p < t_1) = 1 - P(t_1 < t_p) - P(t_0 > t_p) = 1 - \alpha \tag{7}$$

Where probability α can be specified as $\alpha = 0.01, 0.05$ or 0.1 .

From Eq. (5)

$$\begin{aligned}
P(t_1 < t_p) &= P(\bar{X} + K_1 S < \mu + u_p \sigma) \\
&= P\left(\frac{\bar{X} - \mu - u_p \sigma}{S} < -K_1\right) \\
&= P\left(\frac{\sqrt{n}(\bar{X} - \mu)/\sigma - \sqrt{n}u_p}{\sqrt{S^2/\sigma^2}} < -\sqrt{n}K_1\right)
\end{aligned} \tag{8}$$

The statistic $\sqrt{n}(\bar{X} - \mu)/\sigma$ and $(n-1)S^2/\sigma^2$ are independent individually, and

$$\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1) \quad (n-1)S^2/\sigma^2 \sim \chi^2(n-1) \tag{9}$$

Therefore

$$\frac{\sqrt{n}(\bar{X} - \mu)/\sigma - \sqrt{n}u_p}{\sqrt{S^2/\sigma^2}} = \frac{N(0,1) + \delta}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t_{n-1,\delta} \tag{10}$$

Where $t_{n-1,\delta}$ is non-centre t -distribution function with $n-1$ degrees of freedom and non-centre parameter $\delta = -\sqrt{n}u_p$, and u_p is the lower quantile with specified probability p of standard normal distribution.

From Eq. (8) and (10)

$$\begin{aligned}
P(t_1 < t_p) &= P\left(\frac{\sqrt{n}(\bar{X} - \mu)/\sigma - \sqrt{n}u_p}{\sqrt{S^2/\sigma^2}} < -\sqrt{n}K_1\right) \\
&= P(t_{n-1,\delta} < -\sqrt{n}K_1) = \alpha/2
\end{aligned} \tag{11}$$

Then

$$-\sqrt{n}K_1 = t_{n-1,\delta}(\alpha/2) \tag{12}$$

Where $t_{n-1,\delta}(\alpha/2)$ is the lower quantile of the non-centre t -distribution function with $n-1$ degrees of freedom and non-centre parameter $\delta = -\sqrt{n}u_p$ with the given probability $\alpha/2$.

From

$$P(t_0 > t_p) = P\left(\frac{\sqrt{n}(\bar{X} - \mu)/\sigma - \sqrt{n}u_p}{\sqrt{S^2/\sigma^2}} > -\sqrt{n}K_0\right) = \alpha/2 \tag{13}$$

Then

$$-\sqrt{n}K_0 = t_{n-1,\delta}(1-\alpha/2) \tag{14}$$

Where $t_{n-1,\delta}(1-\alpha/2)$ is the lower quantile of the non-centre t -distribution function with $n-1$ degrees of freedom and non-centre parameter $\delta = -\sqrt{n}u_p$ with the given probability value $1-\alpha/2$.

Therefore, when the limited samples value of the maintenance time is given, the maximum maintenance time with the given probability p will fall into the confidence interval $[t_0, t_1]$ with the probability $1-\alpha$.

$$\left[\bar{X} - t_{n-1,\delta}(1-\alpha/2) \frac{S}{\sqrt{n}}, \bar{X} - t_{n-1,\delta}(\alpha/2) \frac{S}{\sqrt{n}} \right] \tag{15}$$

The lower bound t_0 and the upper bound t_1 are as follows:

$$t_0 = \bar{X} - t_{n-1,\delta}(1-\alpha/2) \frac{S}{\sqrt{n}} \quad t_1 = \bar{X} - t_{n-1,\delta}(\alpha/2) \frac{S}{\sqrt{n}} \quad (16)$$

The length of the confidence interval is

$$d = t_1 - t_0 = [t_{n-1,\delta}(1-\alpha/2) - t_{n-1,\delta}(\alpha/2)] \frac{S}{\sqrt{n}} \quad (17)$$

Then

$$d/S = (t_1 - t_0)/S = [t_{n-1,\delta}(1-\alpha/2) - t_{n-1,\delta}(\alpha/2)] \frac{1}{\sqrt{n}} \quad (18)$$

Eq. (18) implies that the length of the confidence interval d/S has relationship with the given sample number n , the probability p and the $1-\alpha$ degree of confidence.

From Eq. (9)

$$E(\bar{X}) = \mu \quad E(S) = \frac{\Gamma(n/2)}{\sqrt{(n-1)/2} \Gamma[(n-1)/2]} \sigma \quad (19)$$

Where the function to calculate sample mean is $E(\cdot)$, $\Gamma(\cdot)$ is the Γ function.

Let $\hat{t}_p = \bar{X} + KS$ ($K>0$ and is constant) be the unbiased estimator value of the maximum maintenance time $t_p = \mu + u_p \sigma$, namely $E(\hat{t}_p) = t_p$, then

$$\hat{t}_p = \bar{X} + \frac{\sqrt{(n-1)/2} \Gamma[(n-1)/2]}{\Gamma(n/2)} u_p S \quad (20)$$

4 Calculations and Analysis

When the probability $p=0.5$ is given, $u_p = 0$ ($\delta = 0$), $t_p = \mu$, Eq. (15) turn to the mean estimate of the maintenance time.

The calculation method for the quantile $t_{n-1,\delta}(\alpha/2)$ and $t_{n-1,\delta}(1-\alpha/2)$ of the non-centre t -distribution function is introduced in the reference [3] and [4].

Shown in the Fig.2, as the sample number n decrease, the un-dimensional interval length d/S increases. Usually when the sample number n is greater than 10, the preferable interval estimation is obtained. If the sample number is too small, the interval estimation is too large to be useful in the engineering.

Shown in the Fig.3, as the given probability value p decrease, the un-dimensional interval length d/S increases. It is appropriate to specify $p=0.9\sim 0.95$ in the interval estimation cases.

Shown in the Fig.4, as the degree of significant α increases, the un-dimensional interval length d/S decrease. It is appropriate to specify $\alpha=0.02\sim 0.1$ in the interval estimation cases.

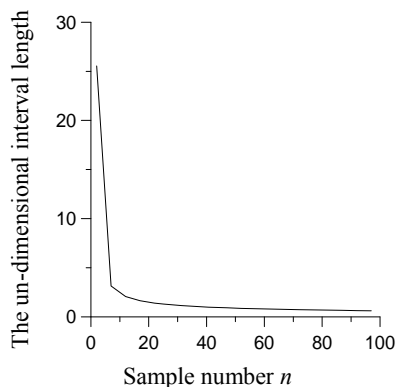


Figure 2 change of the un-dimensional interval length

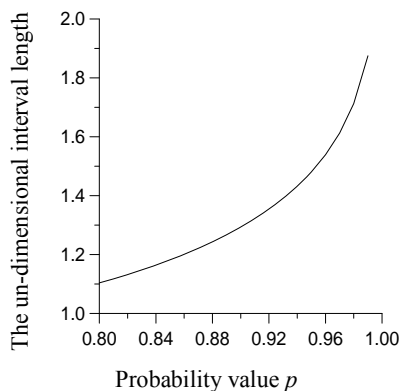
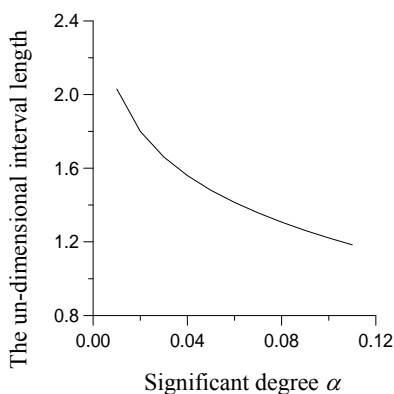


Figure 3 change of the un-dimensional interval length

Figure 4 change of the un-dimensional interval length d/S

5 Application Examples

Some item has the normal distribution. In the maintenance test, the 12 sampling sample value are as follows (unit: h): 190.6, 195.8, 193.0, 218.0, 206.3, 203.5, 185.3, 207.3, 175.1, 194.4, 190.7, 209.8. Find the interval estimation and the unbiased estimation of the maximum maintenance time with the degree of significant $\alpha=0.1$ and probability $p=0.95$.

Solution. The mean and standard deviation of the maintenance time sample are as follows:

$$\bar{X} = 197.483333\text{h} \quad S = 11.901935\text{h}$$

The quantile of the standard normal distribution $u_{0.95} = 1.644854$ with the given probability $p=0.95$. The non-centre parameter δ of the non-centre t -distribution is

$$\delta = -\sqrt{nu_p} = -\sqrt{12}u_{0.95} = -5.697940$$

With the given significant degree $\alpha=0.1$, the quantile $t_{n-1,\delta}(\alpha/2)$ and $t_{n-1,\delta}(1-\alpha/2)$ of the non-centre t -distribution function are

$$t_{11,\delta}(0.05) = -9.478968 \quad t_{11,\delta}(0.95) = -3.680497$$

The lower bound t_0 and the upper bound t_1 of the confidence interval are

$$t_0 = \bar{X} - t_{11,\delta}(0.95) \frac{S}{\sqrt{12}} = 210.128756h$$
$$t_1 = \bar{X} - t_{11,\delta}(0.05) \frac{S}{\sqrt{12}} = 230.051103h$$

Therefore, the interval estimation of the maximum maintenance time with the significant degree $\alpha=0.1$ and the probability $p=0.95$ is $[210.128756, 230.051103]h$

The unbiased estimation of the maximum maintenance time with the probability $p=0.95$ is $t_{0.95} = 217.509480h$.

References

- [1] Zhang Yuzhu, Hu Ziwei, Cao Shimin, Liu Jianguo, Maintenance Validate Test and Evaluation Principles, National Defense Industry Press, 2006, 110-119.
- [2] Zhang Guodong, Lu Tingxiao, Tu Qingci, Analysis and Design of System Reliability and Maintenance, Beijing, BUAA Press, 1990, 264-270.
- [3] Jin Xing, Hong Yanji, Wen Ming, Li Junmei, the application of non-center t-distribution function in reliability lower-bound calculations, Journal of Weapon and Industry, 2003, 24(1):82-84.
- [4] Jin Xing, Hong Yanji, Shen Huairong, Zhang Zheng, Reliability Data Calculation and Application, Beijing, National Defense Industry Press, 2003, 90-94.

Jin Xing, 1962.10, male, Doctor, Longjing of Jilin Province, professor, major in Weapon Reliability and Safety, and System Failure Diagnosis Research.

E-mail: luhai1982@gmail.com

第 3 部分

通信理论与技术

一种认知无线电频谱感知与接入的联合设计方案

丛 容 吴迎笑

(南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘 要: 为更好地利用频谱感知的结果, 实现次用户接入频谱空穴, 本文提出了一种认知无线电频谱感知与频谱接入的联合设计方案。在保证主次用户碰撞率不超过一定门限的条件下, 通过联合选择最佳的感知时间, 发送包长和随机接入概率以最大化网络吞吐量。仿真结果表明物理层感知和 MAC 层接入的跨层联合设计比非跨层设计能更有效地提高网络吞吐量。

关键词: 认知无线电; 频谱感知; 频谱接入; 跨层设计

A Joint Design Approach of Spectrum Sensing and Access in Cognitive Radios

Abstract: In order to allow the secondary users to access the unused spectrum based on spectrum sensing results, a novel joint design approach of spectrum sensing and access in cognitive radios is proposed. In this paper, we design sensing duration, packet length and random access to maximize the secondary network throughput under primary constraints of collision probability. Simulation results indicate that the joint PHY-MAC design performs better than the layered approach.

Keywords: cognitive radio; spectrum sensing; spectrum access; cross-layer

1 引言

随着无线业务的蓬勃发展, 人们更加清楚地认识到无线频谱资源的重要性。传统的频谱分配机制是固定的, 授权用户所占用的频谱不允许其他用户使用。美国联邦通信委员会(FCC)的一份调查报告指出, 现存的频谱授权机制存在大量授权频谱限制, 频谱利用率仅在 15%~85% 之间^[1]。如何进一步提高频谱利用率, 实现频谱资源的动态管理和利用是下一代无线通信亟待解决的问题, 认知无线电技术(Cognitive Radio)作为新的解决方案应运而生^[2]。

在认知无线网络中, 机会频谱接入(Opportunistic Spectrum Access)技术的目标是向次用户(Secondary User)开放授权频谱, 充分利用瞬时频谱空穴。为避免对主用户(Primary User)进行干扰, 次用户必须先进行频谱感知, 发现空闲信道后才能接入进行传输。因此研究认知

无线电环境下的频谱感知和频谱接入尤为重要。Zhao Qing 等在文献[3]中提出用 POMDP 理论研究物理 MAC 跨层情况下的频谱感知和接入策略。文献[4]中提出了基于两种不同频谱感知策略下的认知 MAC 协议，作者运用马氏链模型和 M/GY/1 排队模型分析协议。文献[5]比较了不同机制下的三种动态频谱接入策略，但是跟[4]一样，都没有考虑感知错误的存在以及由此带来对接入的影响。

频谱接入和频谱感知不仅仅只是时间上相互接续的两个部分，它们实际上是相互约束和相互影响的过程。本文的主要工作是物理层感知和 MAC 层接入的跨层联合设计，讨论了在给定主用户足够保护的条件下，如何最大化次用户网络的吞吐量。

2 系统模型

信道根据是否被主用户占用分为繁忙和空闲两个状态，将主用户占用信道看做是两状态时不变一阶马尔科夫链^[6]，如图 1。假设信道空闲和繁忙持续时间服从参数为 v 和 u 的指数分布，这在很多业务过程是满足的，其概率密度函数分别为 $f_v(t) = v \exp(-vt)$ 与 $f_u(t) = u \exp(-ut)$ ，空闲和繁忙的平稳分布分别是 $\frac{u}{u+v}$ 和 $\frac{v}{u+v}$ 。

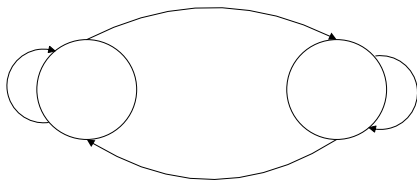


图 1 两状态马尔科夫链信道占用模型

考虑存在一个主用户， n 个次用户的认知无线网络。设次用户的感知时间是 τ ， L_s 是表征次用户进行一次传输所用时间（包长）的随机变量，假设 L_s 是服从均值为 l_s 的指数分布，概率密度分布函数为 $f_{L_s}(t) = \frac{1}{l_s} \exp(-\frac{t}{l_s})$ 。本文约定包长小于平均空闲时间，即 $l_s \leq \frac{1}{v}$ 。

采用能量检测法进行频谱感知，以判断主用户是否存在。检测概率和虚警概率是频谱感知中关键的两个参数。检测概率定义为当主用户占用信道，感知结果检出信道繁忙的概率。虚警概率定义为信道空闲时，错误检测为信道繁忙的概率，分别用 P_d 和 P_f 表示。这里考虑主用户信号是零均值复 PSK 调制信号，次用户接收机处的噪声是服从零均值循环平稳复高斯分布的信号，噪声和主用户信号相互独立，此时虚警概率可以通过下式计算^[7]

$$P_f(\tau) = Q(\sqrt{2\gamma+1}Q^{-1}(P_d) + \gamma\sqrt{\tau f_s}) \tag{1}$$

其中 $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$ ， f_s 为采样频率， γ 为次用户发射机处检测到的信噪比。

传统的信道碰撞模型中，一旦检测到信道是空闲的，多个次用户并发地请求接入同一信道资源。为竞争时减小冲突，本文采用“p-持续”接入机制，即一旦检测到信道是空闲的，所有次用户以概率 p ($0 \leq p \leq 1$) 进行接入。假设有 n 个次用户竞争这一空闲信道，则次用户

能在该信道成功传输的概率为:

$$U = np(1-p)^{n-1} \quad (2)$$

至少有一个次用户传输的概率为:

$$U' = 1 - (1-p)^n \quad (3)$$

不仅次用户间会因为争用信道发生碰撞,主用户与次用户也会发生碰撞。联合设计考虑的是如何选择最佳的感知与接入相关参数以最大化次用户网络的平均吞吐量 $E[Cn]$,同时保证对主用户的干扰在一个可以容忍的范围内。实际次用户相互通信时,还要考虑包头开销的影响,但在此忽略包头因素,其对下面的讨论影响不大。为了保护主用户,我们约定主用户观测碰撞率 D_p 有一个规定的上界 α 。因此本文设计目标如下式:

$$\begin{aligned} & \max E[Cn] \\ & s.t. \quad D_p \leq \alpha \end{aligned} \quad (4)$$

3 感知接入联合设计方案

1) 约束条件

碰撞率是对主用户保护的一个评价参量,定义 D_p 和 D_s 分别为主用户和次用户角度观测的碰撞率^[5]:

$$D_p = \lim_{T \rightarrow \infty} \frac{[0,T] \text{时段内碰撞次数}}{[0,T] \text{时段内信道繁忙次数}} \quad (5)$$

$$D_s = \lim_{T \rightarrow \infty} \frac{[0,T] \text{时段内碰撞次数}}{[0,T] \text{时段内数据包传输次数}} \quad (6)$$

次用户发送的包长是服从均值为 l_s 指数分布的变量,信道空闲时间的概率密度函数为 $f_v(t) = v \exp(-vt)$,则次用户观测碰撞率为:

$$D_s = \int_0^{\infty} f_{l_s}(t) dt \int_0^t v e^{-vx} dx = 1 - \frac{1}{vl_s + 1} \quad (7)$$

在 $[0,T]$ 时段内数据包传输次数表示为 $\frac{T}{l_s + \tau}$,信道繁忙次数为 $\frac{T}{\frac{1}{v} + \frac{1}{u}}$ 。一旦检测到信道是

空闲的, n 个次用户都以概率 p 争用接入, (5) 式中的主用户观测碰撞率由两部分组成:

第一部分是:因信道空闲时,次用户进行接入传输时主用户突然回来引发的碰撞。 $X_1 = p(1 - P_f)$ 表示实际信道空闲时每个次用户的争用接入概率,则这部分碰撞率为:

$$D_{p1} = \lim_{T \rightarrow \infty} \frac{\frac{u}{v+u} \left(1 - \frac{1}{vl_s + 1}\right) \frac{T}{l_s + \tau}}{\frac{T}{\frac{1}{v} + \frac{1}{u}}} (1 - (1 - X_1)^n) \quad (8)$$

第二部分是:信道繁忙时,因为存在漏检概率 $1 - P_d$,次用户误以为信道空闲而接入引起冲突, $X_2 = p(1 - P_d)$ 表示实际信道繁忙时每个次用户的争用接入概率,这部分碰撞率为:

$$D_{p2} = \lim_{T \rightarrow \infty} \frac{\frac{v}{v+u} \left(1 - \frac{1}{vl_s+1}\right) \frac{T}{l_s+\tau}}{\frac{1}{v} + \frac{1}{u}} (1 - (1 - X_2)^n) \quad (9)$$

假设次用户数服从均值为 m 的泊松分布，即 $f(n) = \frac{e^{-m} m^n}{n!}$, $n=0,1,2,\dots$ ，那么平均主用户观测碰撞率为：

$$D_p = \sum_n (D_{p1} + D_{p2}) f(n) = \sum_n D_{p1} f(n) + D_{p2} f(n) \quad (10)$$

2) 目标函数

在间隔 $\tau + l_s$ 内传输时长为 l_s ，每次传输过程不发生碰撞的概率为 $1 - D_s = \frac{1}{vl_s+1}$ ，假设信道容量是 1，那么吞吐量数值上就等于次用户成功传输的时长占信道空闲时长的比重。感知结果不完全正确时的 MAC 层吞吐量由两部分组成：一部分是信道空闲时频谱感知结果是主用户不在时进行接入的吞吐量：

$$\frac{1}{vl_s+1} \frac{l_s}{l_s+\tau} \sum_{n=0}^{\infty} n X_1 (1 - X_1)^{n-1} f(n) \quad (11)$$

还有一部分是信道繁忙时因为漏检，判断主用户不在进行接入产生的吞吐量：

$$\frac{1}{vl_s+1} \frac{l_s}{l_s+\tau} \sum_{n=0}^{\infty} n X_2 (1 - X_2)^{n-1} f(n) \quad (12)$$

结合上面两式，我们可以得到平均吞吐量的表达式：

$$E[Cn] = \frac{u}{u+v} \frac{1}{vl_s+1} \frac{l_s}{l_s+\tau} m X_1 e^{-mX_1} + \frac{v}{u+v} \frac{1}{vl_s+1} \frac{l_s}{l_s+\tau} m X_2 e^{-mX_2} \quad (13)$$

3) 跨层联合设计

传统的设计最大化网络吞吐量将频谱感知与频谱接入分开单独考虑。假设系统规定检测概率 P_d 为一个定值 $\overline{P_d}$ ，虚警概率 P_f 有一个上界 $\overline{P_f}$ 。文献[7]中指出，如果信道繁忙时间比例小于 15%， $P_d=90\%$ ，此时式 (10) 和式 (13) 右边第二项吞吐量可以忽略不计。在上述系统模型中，如果感知结果已知，那问题归结为：感知时间 τ 为定值 $\overline{\tau}$ ，由此虚警概率可由 (1) 式决定，保证此时 $P_f \leq \overline{P_f}$ 。接入的目标是选择合适传输时间 l_s 和接入概率得到非跨层情况下最大吞吐量。单独考虑接入的最优化问题数学表示如下：

$$\begin{aligned} \arg \max_{l_s, p} E[Cn] &= \frac{u}{u+v} \frac{1}{vl_s+1} \frac{l_s}{l_s+\tau} m X_1 e^{-mX_1} \\ s.t. \quad D_p &= \sum_n D_{p1} f(n) = \frac{1}{vl_s+1} (1 - e^{-mX_1}) \leq \alpha, \quad P_d = \overline{P_d} \end{aligned} \quad (14)$$

频谱感知是在物理层上执行，而随机接入属于 MAC 层。为此本文提出采用跨层的思想联合设计感知和接入：系统在保证主用户观测的碰撞率不超过一定门限的条件下，通过联合选择最佳的感知时间，发送包长和随机接入概率三个参数以最大化网络吞吐量。用数学表达式表示如下：

$$\arg \max_{\tau, l_s, p} E[Cn] = \frac{u}{u+v} \frac{1}{vl_s+1} \frac{l_s}{l_s+\tau} mX_1 e^{-mX_1} \tag{15}$$

$$\begin{aligned} s.t. \quad D_p &= \sum_n D_{p1} f(n) = \frac{1}{vl_s+1} (1-e^{-mX_1}) \leq \alpha \\ P_d &= \overline{P_d}, \quad P_f \leq \overline{P_f} \end{aligned}$$

(14) 与 (15) 均可通过最优化理论中基于梯度算法求解,限于篇幅,这里不再详细展开。

4 仿真实验

参数的设定参照 IEEE 802.22 无线区域网标准^[8]。假定信道带宽为 B=6MHz.主用户网络平均繁忙时间为 100ms,非跨层策略中感知时间 $\tau=1\text{ms}$ 。采用用能量检测法检测主用户信号出现与否,采样频率为 $f_s=8B/7$,检测概率 $\overline{P_d}=90\%$,最大虚警概率 $\overline{P_f}=90\%$,碰撞率上界 $\alpha=0.1$ 。

如图 2 所示,将次用户数均值 m 固定在 0.4,比较主用户平均空闲时间 v 从 20ms 到 50ms 变化时,采用跨层和非跨层机制下的次用户最大可达吞吐量取值情况。从图中可见,随着主用户空闲时间的增大,两种机制下的次用户吞吐量都会变大。这是因为主用户空闲概率增大带来次用户可接入的频谱空穴变多,次用户吞吐量增加。另外从图中可以看出,基于跨层的优化方案获得的吞吐量约为非跨层方案的两倍。原因在于跨层的优化方案将影响感知性能的参数也考虑进来,实现了最优的动态频谱接入。

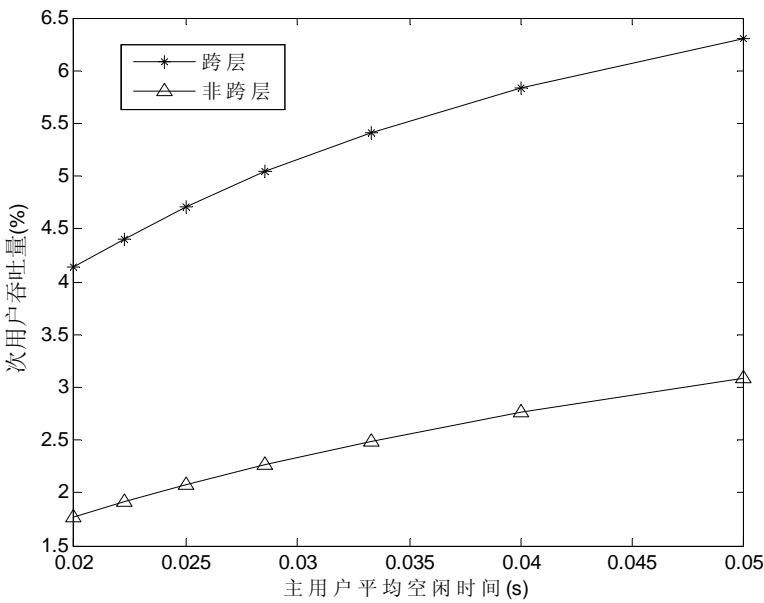


图 2 基于主用户平均空闲时间的两种方案的比较

图 3 是在需要通信的次用户数 m 不断增加的情况下对两种机制所进行的比较。主用户网络平均忙闲时间分别为 100ms 和 10ms,从图中可见,跨层机制较之于非跨层机制,在吞吐量方面始终具有优越性。并且从图中还可以观察到, $m<1$ 时吞吐量随次用户增多而不断增加。

但当 $m>1$ 时, 两者的最大吞吐量都是一个定值不再增加。

如图 4 所示, $m>1$ 时, 两种方案的争用概率 p 都会持续下降, 这是因为当次用户数到达一定量时, 要满足碰撞率的限制, 必须减小争用概率。这也是造成图 3 中 $m>1$ 后吞吐量不再增加的原因。

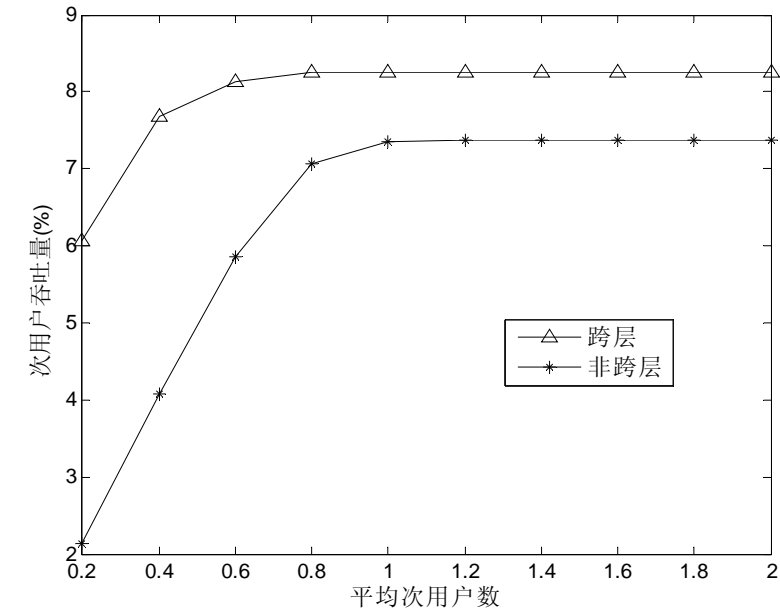


图 3 基于平均次用户数的两种方案的吞吐量比较

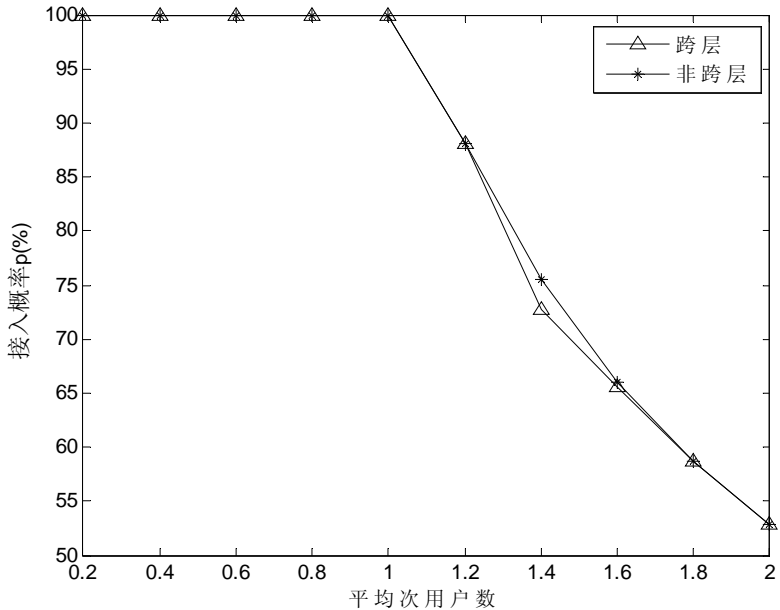


图 4 基于平均次用户数的两种方案接入概率比较

5 总结

认知无线通信系统设计需要考虑频谱感知和频谱接入的优化与折衷, 本文提出了物理层感知和 MAC 层接入的跨层联合设计。通过选择最佳的感知时长, 平均包长以及接入概率来联合设计频谱感知与频谱接入, 将碰撞率控制在实际系统所能忍受的最大范围内并最大化网络吞吐量。通过仿真可以看出, 跨层联合设计相比于非跨层可以获得较大的吞吐量增益。这方面研究对 MAC 层协议设计优化有着重要的意义。

参 考 文 献

- [1] FCC. Spectrum Policy Task Force Report, ET Docket No.02-155[EB/OL]. Nov.02.2002.
- [2] Joseph Mitola, Gerald Q. Maguire, Jr. Cognitive Radio: Making software radio more personal[J]. IEEE Personal Communication, Vol.6, No.4, Aug.1999: 13-18.
- [3] Qing Zhao, Lang Tong, Swami A. Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework[J]. IEEE J. Sel. Areas Commun., Vol.25, No.3, Apr.2007: 589-600.
- [4] Hang Su, Xi Zhang. Cross-Layer Based Opportunistic MAC Protocols for QoS Provisionings Over Cognitive Radio Wireless Networks[J]. IEEE Journal on Selected Areas in Communications, Vol.26, Issue 1, Jan.2008: 118-129.
- [5] Senhua Huang, Xin Liu, Zhi Ding, Opportunistic Spectrum Access in Cognitive Radio Networks[C]. INFOCOM 2008: IEEE 27th Conference on Computer Communications, Apr.2008: 1427-1435.
- [6] Zhiyao Ma, Zhigang Cao, Wei Chen. A Fair Opportunistic Spectrum Access(FOSA) Scheme in Distributed Cognitive Radio Networks[C]. ICC'08: IEEE International Conference on Communications, May.2008: 4054-4058.
- [7] Yingchang Liang, Yonghong Zeng, Edward C.Y.Peh, Anh Tuan Hoang. Sensing-Throughput Tradeoff for Cognitive Radio Networks[J]. IEEE Transactions on Wireless Communications, Vol.7, No. 4, Apr.2008: 1326-1337.
- [8] IEEE 802.22 Wireless RAN. IEEE 802.22-05/0007r46, Functional requirements for the 802.22 WRAN standard[S]. Oct.2005.

作者简介

丛容(1987—), 女, 江苏南通人, 南京邮电大学信号与信息处理专业硕士研究生, 目前主要研究方向为认知无线电。

吴迎笑(1980—), 女, 浙江武义人, 南京邮电大学信号处理与传输研究院博士研究生, 主要研究方向为无线通信与网络信号处理。

多用户检测算法及其simulink仿真研究

任大山 龙 昕 杨明华

(云南大学信息学院, 云南昆明, 650091)

摘 要: 作为 3G 技术人增强技术之一的多用户检测由于能很好地减少多址干扰和解决远近效应问题, 显著提高了系统容量。使其愈来愈受到学术界、产业界的重视。通过对现有多用户检测技术分析研究的基础上采用 Simulink 全新的动态建模形式对其性能进行仿真分析, 使其仿真模型的建立更加简洁且模型更具可读性和可扩展性, 对多用户检测的理论及其教学研究有实践及其应用价值。

关键字: MAI; 远近效应; CDMA; 多用户联合检测; Simulink。

Research on Multiuser Detection Arithmetic and Modeling with Simulink

REN Da-shan LONG Xing YANG Ming-hua

(College of Information, Yunnan University, Kunming (650091), Yunnan, China)

Abstract Multiuser detection as one of enhancement techniques of 3G could decrease the multiple access interference and resolve the near-far effect, improve system capacity. It is taken more seriously by academic and industrial circles. It adopts a novel Simulink dynamic modeling to model the modeling on research on multiuser detection arithmetic. So it makes modeling easy, simple, readable and scalable, It is very valuable for theoretical research and application.

Keywords: MAI; Near-far Effect; CDMA; Multiuser Detection; Simulink

1 概述

2009 年 1 月 7 日, 工业和信息化部为中国移动、中国电信和中国联通发放 3 张第三代移动通信(3G)牌照, 此举标志着我国正式进入 3G 时代。实用的第三代移动通信系统(3G)已经走入我们的生活。由于 CDMA 系统是一个干扰受限系统, 系统中存在多址干扰和远近效应, 这两种因素是限制 CDMA 系统容量的主要因素。多用户检测技术作为 TD-SCDMA、cdma2000、WCDMA 的增强性技术之一, 在提高系统容量方面起着举足轻重的作用。要描述动态系统的特性, 传统的建模方法是先对系统的输入信号和输出信号进行分析, 得到它们的系统方程, 然后编写程序进行仿真。这种仿真方法有两个缺点。首先是不够直观, 缺乏足够的人机交互。由于所有的输入信号和输出信号都被抽象成数值之间的关系, 仿真表现为一种计算过程, 因

此难以对仿真的过程进行控制，也难以对仿真的输出数据进行直观的描述和分析。另外，这种方法缺乏系统性，尤其是在对复杂系统的处理过程中，难以采用模块化方法，从而降低了仿真程序的可读性和可扩展性^[1]。

2 多用户检测的发展现状及其发展趋势

多用户检测的思想诞生于 1979 年，Schneider 第一次将多个用户的码字和定时信息联合起来检测单个用户的信息并研究了迫零法^[2]。1983 年 R.Kohno 发表了对多用户干扰消除器(IC)的研究，利用其他用户的已知信息消除 MAI，实现无 MAI 的多用户检测^[3]。后来 Veulu 于 1986 年率先提出了最优多用户检测，这种检测器是一种最大似然估计算法^[4]。这种算法从理论上分析可以达到单用户接收机的性能，也能有效地克服远近效应，但由于此种算法需要的已知量太大，而 CDMA 系统中接收机又往往只知道自己感兴趣的用户的信息，况且此种算法的复杂度会随系统中用户数目的增加而呈指数级增加，由于这些因素的影响，导致这种最优算法在实际工程中基本无法实现。于是人们开始研究各种次优多用户检测器，目的就是保证以工程上可以实现的复杂度使系统性能得到一定的提高。

2.1 单用户模型

首先我们从单用户角度出发，假设一发射机发射通过一无线信道发射一个符号序列为 {b[0],b[1],b[2],...b[M-1]}，此处我们考虑线性调制。则我们可将这个发送信号的波形表示为^[5]：

$$x(t)=\sum_{i=0}^{M-1}b[i]w_i(t) \tag{2-1}$$

其中 $w_i(t)$ 是与第 i 个符号相对应的调制波形，其形式通常可用如下表达式表达：

$$w_i(t)=Ap(t-iT)e^{j(w_c t+\phi)} \tag{2-2}$$

式 2-2 中 $A>0$ ，相位偏移 $\phi\in(-\pi,\pi)$ ，而 $p(t)$ 是周期为 T 的基带波形，它可以是一个矩形脉冲，也可以是一个升余弦或带限脉冲。特别的当系统是一个直接序列扩频系统时，

$$p(t)=\sum_{j=0}^{N-1}c_j\psi(t-jT_c) \tag{2-3}$$

其中 N 为扩频增益， c_j 是伪随即扩频码 ($c_j\in\{-1,1\}$)，而 $\psi(t)$ 是码片波形。

2.2 多用户检测模型

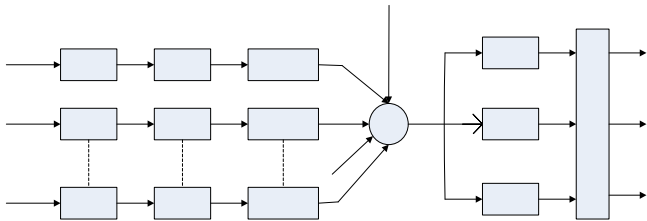


图 2-1 典型的多用户模型

由以上模型我们能将一个在高斯白噪声环境下有 K 个用户的 CDMA 系统的接收信号用如

下的数学表达式表示^[6]:

$$r(t)=\sum_{k=1}^KA_k\sum_{i=0}^{M-1}b_k[i]s_k(t-iT)+n(t) \tag{2-4}$$

倘若现在我们只对接受波形的第一个用户的数据感兴趣，这时我们可以设计这样的一个加权向量 $w_1 \in C^N$,则期望用户的数据比特率就按如下方式解调： $z_1[i]=w_1^Hr[i]$ ，把 2-4 带入得：

$$z_1[i]=A_1(w_1^Hs_1)b_1[i]+\sum_{k=2}^KA_k(w_k^Hs_k)b_k[i]+w_1^Hn[i] \tag{2-5}$$

由 2-5 式中可以看出其第一项就是包含我们所需的第一个用户的信息，第二项包含非期望用户的信息比特，也就是我们常说的 MAI（多址干扰），下面我们就来讨论一下一种线性多用户检测及其优化算法。

3 解相关用户联合检测

首先看一下线性解相关检测器，其思想是假设能有一个 w_1 使得 $w_1^Hs_1=1$ 而 $w_k^Hs_k=0$ ，则 2-5 中既能完全检测出我们所期望的用户的信息又能完全抑制多址干扰。在此定义用户 1 的线性解相关检测器 d_1^H （ $w_1 @ d_1$ ）满足一下条件^[7]（式 3-1，3-2）：

$$d_1^Hs_1=1 \tag{3-1}$$

$$d_1^Hs_k=0 \tag{3-2}$$

由以上两个关系式我们可以得出一个线性解相关检测器的表达式（式 3-3）：

$$d_1=SR^{-1}e_1 \tag{3-3}$$

式中 $S @ [s_1,s_2,...,s_k]$ ，此时假设用户的特征序列是线性独立的，则 $\text{rank}(S)=K$ ，所以我们可以假设有一个可逆矩阵使得 $R @ S^H S ; e_k$ 表示除了第 k 个元素为 1 以外，其他元素均为零的 K 维向量。此时线性相关检测器的输出：

$$z_1[i]=d_1^Hr[i]=A_1b_1[i]+d_1^Hn[i] \tag{3-4}$$

由 3-4 式可以看出这种检测器能很好的检测出期望用户数据，且完全消除 MAI，根据柯西-许瓦次不等式有： $Pd_1^2.Ps_1^2 \geq Pd_1^Hs_1P^2$ ，由于 $Ps_1P=1$ ， $d_1^Hs_1=1$ 可以得出 $d_1 \geq 1$ ，因而促使噪声功率增强。在信噪比较大的环境中对期望用户的检测影响不是很大，但当信噪比较小时，期望信号就会被噪声信号所淹没。从而不能正确检测出期望用户的数据。

4 线性最小均方误差多用户联合检测（MMSE）

由上面的解相关检测我们可以看出其是一种以牺牲信噪比为代价而消除 MAI 的方法，在信噪比较低的情况下检测结果将差强人意，为此人们希望寻找到一种较为折中的算法，这样 MMSE 多用户联合检测便应运而生了。其思想是使从检测器中输出的多址干扰和输出噪声对数据检测总的的影响最小^[8]。由此我们可以把求解某个（第一个）用户的线性 MMSE 检测器转化为对下面这个问题的求解：

$$m_1 = \arg \min_{w \in \mathcal{N}} E\{\|A_1 b_1[i] - w^H r[i]\|^2\} \quad (4-1)$$

再次我们把 m_1 定义为 MMSE 多用户检测的检测器，即： $w_1 @ m_1$ 。倘若我们把 K 个用户的增益列成一个三角矩阵，即定义 $|A| @ diag(|A_1|, |A_2|, |A_3|, ..., |A_k|)$ ，我们现在可以得出用户 1 的线性 MMSE^[9]：

$$m_1 = S(R + \sigma^2 |A|^{-2})^{-1} e_1 \quad (4-2)$$

至此可以得出线性 MMSE 检测的表达式，如 4-3 所示。

$$z[i] @ m_1^H r[i] = A_1(m_1^H s_1) b_1[i] + \sum_{k=2}^K A_k(m_k^H s_k) b_k[i] + v[i] \quad (4-3)$$

其中， $v_1[i] @ m_1^H n[i] : N_c(0, \sigma^2 \|m_1\|^2)$ ，我们仔细观察式 4-3 会发现，在 MMSE 检测器的输出中，不仅包含一定的 MAI 冗余信息，还存在一定的噪声干扰。而对于线性解相关检测器和 MMSE 的性能我们将会在下一节中进行讨论。

5 simulink 动态仿真及其建模

本文利用 Simulink 平台，模拟了一个小区内十个用户的码元发送，扩频，接收，解扩，判决的 CDMA 通信基本过程，实际仿真模型图 5-1 所示。

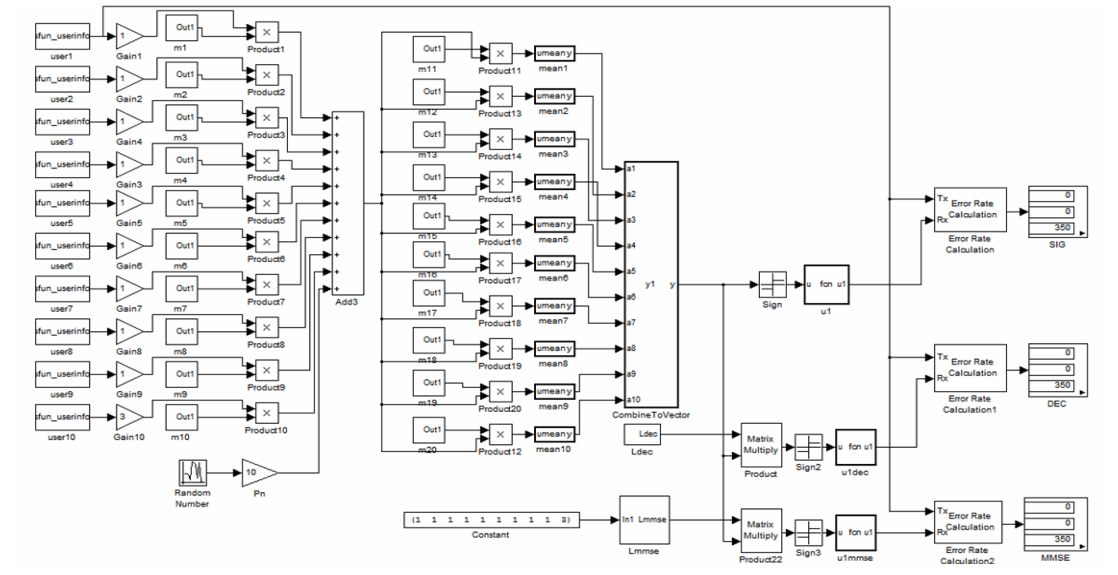


图 5-1 CDMA 多用检测的 Simulink 仿真模型

本平台共包括用户信号生成模块、m 序列发生器模块、Add 模块、积分判决、解相关多用户检测和最小均方误差多用户检测的线性算子用自定义模块等几个模块。其中信号生成模块由自己编写的一个 S 函数产生 10 个用户的随机信源码，其值是离散的取 ±1，分别进过经调制和加入高斯白噪声，m 序列则运用 Embedded MATLAB Function block 编写了一个 m 序列发生器模块，其输出为一个 P=31 的 m 序列。而模型中的远近效应由增益模块的参数设置来实现。整个模型中的模块及其函数的具体实现细节在此不做过多的叙述，接下来看一下运行

的仿真结果。如图 5-2 所示。

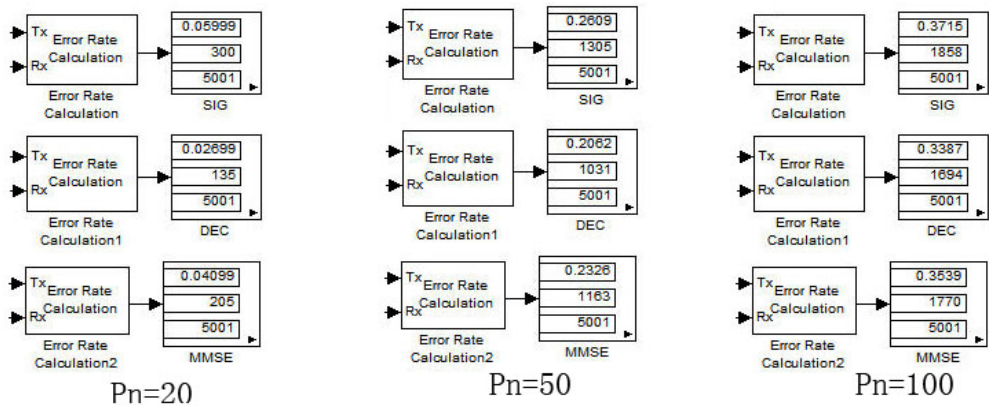


图 5-2 三种多用户检测算法误码性能的比较仿真

上图中仿真的是在三种不同的噪声功率输入情况下的情况，本图自上而下分别表示传统的单用户检测、多用户检测和最小均方误差检测的误码性能，每个显示结果中自上而下，分别表示误码率、错误比特以及仿真时间。由上图可以看出在同等信噪比情况下多用户检测和最小均方误差检测的抗误码性能要高于传统的单用户检测，由于存在一定的误差，所以解相关检测器和 MMSE 检测器的抗误码性能相仿（实际上在噪声功率较高时抗误码性能应该是后者优于前者）而且随着信噪比的不断降低误码率随之提高。

6 结论

本文通过对多用户检测算法进行详细的推导研究，并且采用非常规的动态仿真模式，建立 CDMA 中三种检测器的仿真模型，并通过实际仿真结果验证了其正确性和优越性。从整个仿真过程可以看出，用 Simulink 的动态仿真模式较传统的静态仿真有着较好的可读性和扩展性，本文中只写了单用户、解相关、和 MMSE 三种算法的模型，倘若需要对更多的算法建模，可直接编写子模块与之连接即可，增强了模型的可扩展特性。从仿真结果中来看，在抗误码性能和抗远近效应方面 MMSE 和解相关多用户检测器存在一定的优越性，但解相关检测器会使输出噪声增强，而 MMSE 则是均衡了多址及其噪声对输出信号检测的影响。但总的来说多用户检测在 CDMA 中对消除多址干扰和抗远近效应方面有着举足轻重的意义。

参 考 文 献

[1] 张德丰. MATLAB /Simulink 建模与仿真[M]. 北京：电子工业出版社，2009.

[2] S verdu. Optimum multiuser asymptotic efficiency. IEEE Trans Commun. 34(9):890-897,1986.

[3] R Kohno,H Imai, M Hatori. Cancellation technique of co-channel interference inasynchrns spread spectrum multiple access system[J]. IEICE Trans Commun. 1983,66:20-29.

[4] S Verdu. Minimum of error for asynchronous Gaussian multiple-access channels[J]. IEEE Trans Inform Theory.1986,32:85-96.

[5] 张贤达. 现代信号处理(第二版)[M]. 北京：清华大学出版社，2002.

- [6] DU Y G, CHAN K T . Improved multiuser detectors employing genetic algorithms in a space-time block coded system [J] . EURASIP Journal on Applied Signal Processing, 2004, 2004 (5) : 640 - 648.
- [7] Xiaodong Wang, H. Vincent Poor. Wireless communication systems advanced techniques for signal reception[M]. 北京: 电子工业出版社, 2005.
- [8] 龚秋莎, 杨家玮. 第三代移动通信系统中的多用户检测[J] . 电信快报, 2002 (11) : 21-22.
- [9] 徐大勇, 肖扬. 用于 CDMA 系统的一种改进的 MMSE 多用户检测算法. 中国科技论文在线. <http://www.paper.edu.cn>.

On Design and Simulation of Electrostatic Sensor Used for Measuring Gas-Solid Two-phase Flow

Yu Zhi-gen

(Faculty of Electromechanical Engineering, Huzhou Vocational and Technical College, Huzhou 313000, China)

Abstract: By analyzing the current theory model of electrostatic sensor, we present an improved one that based on point charge. The sensing mechanism is described precisely because influencing factors such as the geometry of sensing element, the tubing insulating material, the geometry of electro-magnetic shielding around are considered synthetically in this model. A reasonable structure is finally got after simulated by analysis software of the finite element, and electrostatic sensor designed according to this model is widely used to measure the density of gas-solid two-phase flow.

Key words: gas-solid two-phase flow, electrostatic sensor, finite element analysis

1 Introduction

The gas-solid two-phase flow is a common logistics mold in production process. Pulverized coal, cement, ore, salts, flour are all conveyed by pipeline in industries of power, building material, metallurgy, environmental sanitation, medicines and grain processing. It becomes more and more important to measure precisely the parameters of gas-solid two-phase flow with the high demands on detecting parameters and control. For instance, it has become an important study to control the pollution sources by detecting and controlling dust emitted from industries. The conventional method to measure the speed and density of gas-solid two-phase flow is to construct a computing model by deriving from experimental analyze and introducing a correction factor^[1]. It has also become a forward subject concerned by scholars at home and abroad because the research on multi-phase flow is constrained by the technology of measuring the gas-solid two-phase or multi-phase flow^[2].

2 An Improved Model of Measuring Electrostatic Sensor Based on Point Charge

Three methods are used to detect the electrostatic particle: triboelectricity, electrodynamics, electrostatic. There are fine differences in expressing the sensing principles. In triboelectricity, particles are charged by friction or their contact or collision with sensing probes. While in

electrodynamics, dynamic characteristics of measuring mechanism are emphasized, and only those moving charged particles affect the sensor's output. The electrostatics emphasizes the characteristics of electrostatic induction and thus is well suitable for the name. Fig.1 shows the system model for measuring parameters of gas-solid two-phase flow that based on particle staticelectricity. Its fundamental principle is that particle staticelectricity is detected by a charge-converted circuit according to the different charging mechanism of the tested objects. Electrical signals are output to measuring instrument after being amplified and filtered. Hence, parameters of gas-solid two-phase flow are detected online.

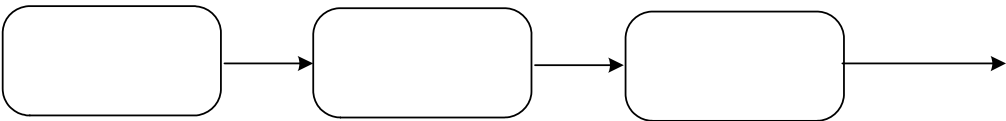


Figure 1 diagram of electrostatic sensor for measuring gas-solid two-phase flow

The current sensing elements has two structures, one is contact (ring-shaped), the other is non-contact (rod-like). The sensor electrodes are required to have good conductive ability ^[3]. The ring-shaped electrostatic sensor is an ideal one used to detect parameters of multi-phase flow in industrial production and will be introduced in the following.

2.1 Research Status of the Ring-shaped Model of Measuring Electrostatic Sensor

The essence of the ring-shaped electrostatic sensor is to employ its interaction with the electrostatic field formed by charged particles. The outputting induced charges and potential vary with the electrostatic field because of the moving particles. It is necessary to build a mathematical model in order to verify the interaction between charged particles and sensors, to understand its sensing mechanism deeply, to study quantitatively the sensing characteristics such as spatial sensitivity, spatial filtering effect, and frequency response, to provide theoretical basis necessary for determining and evaluating its performance, optimizing design, and improving its dynamic performance. Many models are built by scholars such as Gajewski from Technical University of Wroclaw in Poland [4], YanYong ,Cheng from Tesside university in U.K [5]., and Murnane, Woodhead from University of Greenwich [6]. Although these models have been widely applied in researching on measuring parameters of gas-solid two-phase flow, there are massive assumed conditions in building models of dynamic mechanism, and both the spatial filtering characteristic and the frequency bandwidth characteristics are given.

Harsh conditions such as equipotential particles are required in Gajeski's model. In fact, it's impossible for non-conducting particles to be equipotential. Furthermore, the mutual-capacitance and self-capacitance are different for particles in different sensing area. So even for particles with same electricity but in different space, the induced potential in electrodes will change, which is not considered in this model. It's also impossible that in pneumatic conveying system, the particles are steady flow and disperse uniformly in pipeline. Form multiphase fluid mechanics, we know that

even dispersed particles have different concentration in local space. Besides, charges carried by particles are different in every possible way. So the relation between charge's density and induced potential couldn't show the reaction between electrostatic field and electrode. More importantly, Gajeski didn't take into account the sensor's dimension that affects its performance in fact. To sum up, the model could only explain quantitatively the relation between the induced potential and the charges carried by particles or volume charges density. Its output affect by spatial distribution couldn't be analyzed without the particle's sensitivity distribution.

2.2 An Improved Model of Measuring Electrostatic Sensor

Models showed above describe the sensing mechanism in some extent, and provide theoretical basis for analyzing characteristics. There are too many assumed conditions in building the models and only axial length are taken into account. In practice, the electrostatic sensor is composed of sensing elements, insulated tube and electro-magnetic shielding. All of them should be taken into account in analyzing sensor's characteristic or in building mathematical models. Besides, when charged particles flow through electrode, equal but opposite charges are induced in the internal and external surface of the electrode. Charges induced are sum of those produced by particles' charges in the induction zone. So a single particle charge could be thought of as a point charge. The electrostatic sensor model here is built on the basis of point charge, and employs Gajeski's structure. As shown in Figure. 2, it could stand high pressure and is electric insulated because the insulated tube is made up of quartz glass. For modeling conveniently, let's assume that point charges in radial position move axially at a constant speed; ignore the magnetic effect of moving charges; assume that the particle charges that flow through the electrode is saturated, that the transmission pipeline a metal one with good conductivity and is grounded, that the electrode is a mental ring with extremely good conductivity and shunt capacitance is ignored. This model is more suitable for pneumatic transporting pulverized coal, cement and etc.

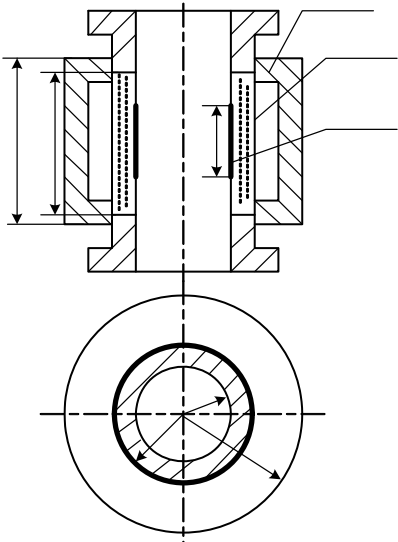


Figure 2. structure model of an improved electrostatic sensor based on point charge

Because the interaction between the field formed by inducing charges and that formed by induced charges, conductor could keep electrostatic equilibrium in very short time (10^{-19} s)^[7]. The interaction between the moving point charges and the electrostatic sensor can be described by electrostatic field. When point charges are in the sensing zone, the electrostatic field satisfies the following Poisson equation and the boundary condition:

$$\begin{cases} \nabla(\varepsilon(x, y, z)\nabla\phi(x, y, z)) = -\rho(x, y, z)\Phi(x, y, z) \\ \Phi(x, y, z)\big|_{(x, y, z) \in \Gamma_p} = 0 \\ \Phi(x, y, z)\big|_{(x, y, z) \in \Gamma_s} = 0 \\ \Phi(x, y, z)\big|_{(x, y, z) \in \Gamma_e} = Cons \end{cases} \quad (1)$$

According to Gauss theorem, density of the induced charges in the electrode's surface is:

$$\sigma(x, y, z) = \hat{D}(x, y, z) = \varepsilon(x, y, z)\hat{E}(x, y, z) = -\varepsilon(x, y, z) \bullet \nabla\phi(x, y, z) \quad (2)$$

Where, $\phi(x, y, z)$ is a field potential with the unit of V; $\rho(x, y, z)$ is a density distribution of the volume charges with the unit of C/m³; $\varepsilon(x, y, z)$ is a dielectric constant distribution with the unit of F/m; Γ_p 、 Γ_s are boundaries formed by pipeline and shielding; Γ_e is an electrode boundary; Cons means that the sensor electrode is equipotential; $\sigma(x, y, z)$ is a density distribution of induced charges in electrode with the unit of C/m²; $D(x, y, z)$ is an electric displacement vector near the electrode inner wall with the unit of C/m², $E(x, y, z)$ is a electric field distribution with the unit of V/m.

The quantity Q of induced charges in the electrode's inner surface with area of S is computed by the Equation (2):

$$Q = \int_S \sigma(x, y, z) \bullet ds \quad (3)$$

From the mathematical model above, we can see that in a sensitive space, under conditions of the known $\varepsilon(x, y, z)$, $\rho(x, y, z)$, and boundary, the quantity Q of induced charges can be computed by Equation (1), (2) and (3). Because of the complex boundary condition and flow in pipeline, it is relatively difficult to parse the mathematical model and the induced charges are computed by finite element simulation.

3 Finite Element Simulation of Sensing Element

With graphic interface, program structure, and interactive graphic processing, ANSYS software can be used for computing structure, fluid, thermal, electromagnetic, and coupled field, which alleviates workloads of modeling, finite element solution, results' analysis and evaluation. For electric field, we can analyze current's conduction, circuits, field-circuit coupling, and electrostatic field. Electrical quantities such as current density, electric field intensity, and capacitance can be solved as well^[7]. ANSYS8.1 software of PC version, together with ANSYS APDL, is used to program sensitivity distribution in sensing field of the electrostatic sensor. And parameters such as structure ones, potential inside the sensing field, induced electric, sensitivity can be analyzed,

simulated, and evaluated.

3.1 Finite Element Analyzing Electrostatic Field of the Sensing Elements in Electrostatic Sensor

When charged particles move in the pipeline, induced charges will be produced in the electrode and the metal pipeline, and there will be an interaction between the electrostatic field formed by induced charges and that formed by charged particles. So, the induced charges distribution in electrode and the induced electricity will be affected by the electrode and pipeline inevitably. Because the interaction between the field formed by inducing charges and that formed by induced charges, conductor could keep electrostatic equilibrium in very short time (10-19s). The field can thought of as an electrostatic one according to the criterion of near steady in physics. At present, there are five methods used to solve the modeling: analytical method, boundary finite element method, finite difference method, boundary finite element method, and Monte Carlo method. Which one is chosen just depends on geometry's complexity and dimension of problems to be solved, and software suitable. Approximate solutions will be got in finite difference method. With the development of electronics and computer technology, the finite difference method has been widely applied in solving problems in electromagnetic field such as electrostatic field, time-varying field, and nonlinear field etc. The best advantage of it is that it fits for problems with complex boundary and complicated media distribution, and is not limited by the boundary shape. Hence, the finite difference method is selected to analyze its sensing characteristics and to design its sensing elements optimally.

Problems in electrostatic field can be turned into not only definite problems of differential equation, but also variational problems in which extreme values are to be solved. These two are equivalent according to Variational Principle. Based on Variational Principle, the finite element method developed form difference scheme. First, boundary value problems expressed in differential equation is turned into variational problems that getting extreme values. Secondly, the continuous field is discretized by field subdividing. A finite element subspace is constructed to turn approximately the variational problems into solving extreme value of multivariate function^[8]. According to Thomason theorem, charge distribution in the surface of charge conductor enables the electrostatic field to have minimum energy. And equation (4) is gotten:

$$\left\{ \begin{array}{l} J_{\varphi} = \iiint_D \frac{\varepsilon(x,y,z)}{2} \left\{ \left[\frac{\partial \varphi(x,y,z)}{\partial x} \right]^2 + \left[\frac{\partial \varphi(x,y,z)}{\partial y} \right]^2 + \left[\frac{\partial \varphi(x,y,z)}{\partial z} \right]^2 - \rho \varphi \right\} dx dy dz = \min \\ \Phi(x,y,z) \Big|_{(x,y,z) \in \Gamma_p} = 0 \\ \Phi(x,y,z) \Big|_{(x,y,z) \in \Gamma_s} = 0 \\ \Phi(x,y,z) \Big|_{(x,y,z) \in \Gamma_e} = Cons \end{array} \right. \quad (4)$$

Where, D is the shielding inner space.

3.2 Finite Element Model of Electrostatic Sensor

Let the axial direction of the electrode be Z-axis, the radial be r-axis, and the circumferential be θ -axis in the global coordinate system. Two electrostatic field distributions that formed by point charges at any site of the infinite grounding cylinder: one is a two-dimension field when point charges at the axis, the other is a three-dimension filed when charges deviate form the axis. The finite element analysis method is quick enough for two-dimension field but slow for three-dimension filed. The three-dimension filed can be converted to two-dimension in conditions of that per point charge is supposed to distribute uniformly at the charge's radial circumference. Both the pipeline itself and constraint formed by liner charge are axisymmetric. Although electrostatic field formed by liner charges is different form that formed by point charges, they are equivalent for induced charges in the electrode because electrostatic fields at the same axial and radial position have the same induced electricity by Superposition Principle. Modeling thus becomes dramatically easy and computing is speeded up.

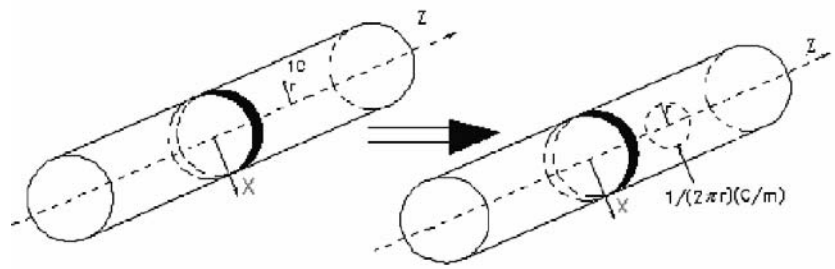


Figure 3 Simplifying three-dimension electrostatic filed to two-dimension one

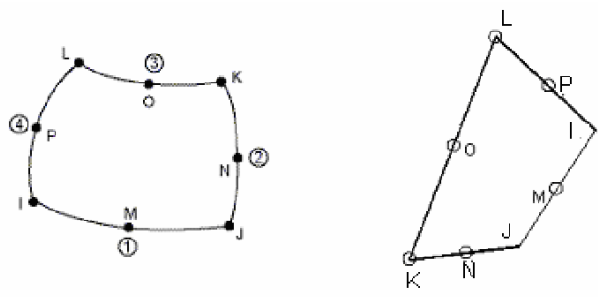


Figure 4 PLANE121 geometry Figure 5 INFINE110's geometry

From the analysis above we know that point charge at any position can be solved according to a two-dimension field. The PLANE121 (shown in Fig.4) which has a two-D quadrilateral with eight nodes or triangle axisymmetrical unit is selected according to the probe's size and shape of and the demanding computing precision. In a far-field area, the INFINE110 (shown in Fig.5) which is an axisymmetrical unit with a two-D quadrilateral eight nodes matches the PLANE121. Besides, the electrostatic sensor is divided into three areas: the inner insulated pipe, the wall of insulated pipe, and the shielding space between the outer pipe and the shielding case. A dense grid is divided

because there is a drastic changing potential near the point charges, electrode, and the shielding case. Fig. 6 shows the axisymmetrical grid division of point charges locate in the center of the axis and Fig. 7 shows that of point charges off the axis center.

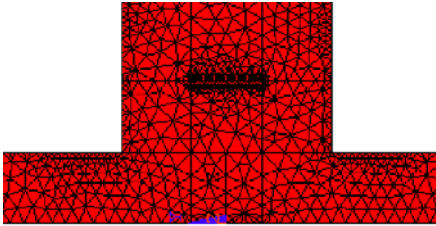


Figure 6 Grid division of the finite element model with point charges in the center of the axis

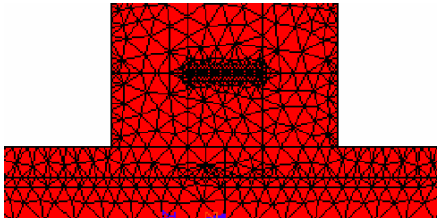


Figure 7 Grid division of the finite element model with point charges off the center of the axis

3.3 Analyzing Results of Simulation Experiment

We draw several useful conclusions based on simulation experiment on the sensing elements:

- 1) A longer axial length (W_e) results in higher sensitivity, a more homogeneous distribution of the section's sensitive field, and a wider sensitive axial space. A too long electrode doesn't enables the sensor to response to the high frequency signal of the fluid space because of the special filtering effect of the electrostatic sensor, while a too low electrode results in a small output, a low signal-to-noise ratio, and a difficult detection.
- 2) The tube-wall's thickness (R_2-R_1) and its relative dielectric constant (ϵ_{ri}). When the thickness satisfies mechanical strength, the thinner, the better. A thick tube-wall lows the absolute sensitivity because it occupies the high sensitive area. However, an increased thickness contributes to the uniformity of the section's sensitivity. An increased relative dielectric constant ϵ_{ri} increases its absolute sensitivity and a non-uniform sensitivity of the section, but almost hasn't any influence on the sensitivity of the axial space.
- 3) Radial size of the shielding case (R_3-R_2) and axial length (L). The radial size of the shielding case has less influence on the sensitivity and its distribution. But the absolute sensitivity of the central section increased with the axial length, which leads to an increasing the average sensitivity and contributes to the uniformity of the section's sensitivity. An increased axial length results in a high sensitivity and a wide sensitive space. That is , an appropriately increased axial length makes circuit design easier.

4 Conclusion

On basis of analyzing the merits and demerits of the electrostatic sensor models, we give an improved model which thinks of the probe made up of sensor electrodes, insulated pipe and electromagnetic shielding as a whole. By ANSYS8.1 (an infinite element analysis software)of PC version and APDL, an infinite element program is composed to computer the sensitivity distribution

of the electrostatic sensor. Feasible scheme is provided as well to modeling and to research on the characteristic of the sensor for measuring two-phase flow.

Reference

- [1] National Natural Science Foundation of China. Investigation Report on Developing Strategy of National Science- - Engineering Thermophysics and Energy Utilization [R]. Beijing: Science Press.1995.
- [2] Y •Yan. Flow Measurement of Particulate Solids in Pipelines [J]. Flow Measurement and Instrumentation, 2000, 11:151.
- [3] Y Yan and J Ma. Measurement of Particulate Velocity under Stack-flow Condition [J]. Meas.Sci. Technol., 2000, 11:59–65.
- [4] Gajewski J.B. Dynamic Effect of Charged Particles on the Measuring Probe Potential [J]. Journal of Electrostatics, 1997, 40&41:437-442.
- [5] Cheng R.A Study of Electrostatic Pulverized Fuel Meters [D]. Ph.D. Thesis, University of Teesside,U.K.1996
- [6] Zushou Zhang, On Magnitude Estimation of Time for Conductors to Reach Electrostatic Equilibrium [J].Physics and Engineering,2003,12:20-21.
- [7] Shuguang Gong, Ansys Operation Command and Parametric Programming [M].Beijing :Mechanical Industry Press ,2004:231.
- [8] S.Matsusaka,H.Masuda .Electrostatics of Particles[J].Journal of Electrostatics,2003,14:143-166.

Brief introduction of the authors:

YU Zhi-gen (1965-), an associate professor in Huzhou Vocational and Technological College, a visiting scholar of Zhejiang University, majoring in researching in electrical power and electron.

协作分集中的移动中继动态选择和切换策略研究

张 鑫 谢显中 雷维嘉

(重庆邮电大学个人通信研究所, 重庆 400065)

摘 要: 在协作无线通信系统中, 中继节点的移动会大大降低系统的性能, 目前这方面的结果很少。本文探讨协作分集中的移动中继选择算法, 在放大转发 (AF) 协作通信模式下, 给出了基于瞬时信道状态信息的功率分配及中继的动态选择策略。针对单中继情况提出了中继切换方案, 针对多中继情况提出了动态剔除、补充中继的方案, 并分析了相应的分集增益及系统容量。通过模拟仿真分析, 该方案能有效的降低中断概率, 提高系统分集增益, 扩大系统容量, 实现良好的整体性能。

关键词: 协作分集; 移动中继; 瞬时信道增益; 选择策略; 切换

Selection and handoff Scheme of mobile Relay in Cooperative Wireless Networks

ZHANG Xin XIE Xian-zhong LEI Wei-jia

(Institute of Personal Communication of Chongqing University of Posts and Telecommunications, Chongqing 400065, P.R. China)

Abstract: The mobility of the relays in cooperative wireless networks will greatly reduce the performance of the system, so far there is not much result on this point. We discuss the relay selection and handoff scheme of mobile relay in cooperative diversity networks and present the power allocation algorithm and relay selection algorithm based on instantaneous channel information under amplify-and-forward (AF) mode. For a single relay we use the handoff scheme and in other situations we choose the relay selection scheme. Furthermore, we analyze its diversity gain and system capacity. Simulation results show that the proposed relay selection and handoff scheme effectively reduce the outage probability, increase its cooperative gain and system capacity.

Keywords: cooperation diversity, mobile relay, instantaneous channel gain, selection scheme, handoff

本文获国家自然科学基金 (60872037)、重庆市自然科学基金 (2008BB2411) 和重庆市教委应用基础研究基金 (KJ080508) 资助。

引言

利用多天线 (MIMO) 技术在不增加系统带宽和天线发射总功率的情况下, 能成倍地提高系统容量, 对系统性能的改善十分显著, 是对抗无线信道中多径衰落的有效手段, 协作分集使单天线的移动终端也可以实现 MIMO 传输, 是目前的研究热点。协作的概念最早由 Sendonaris 等人提出^[1-2], 随后 Laneman 等研究了各种具体的协作通信协议^[3-5]。

协作分集的基本思想是信息的传递过程中, 源节点可以利用一个或者多个中继节点协助通信, 降低中断概率, 扩大系统容量, 带来了协作分集增益。近年来有大量的文献对基于固定中继的协作通信系统的中继选择、功率分配以及性能进行了研究^[6-10]。Zhou Kenan 等^[6]给出了检测转发 (DF) 模式下的一个单个固定中继的选择方案, 在理想功率选择 p_i^* 的前提下得出了理想的资源到中继的距离, 该功率分配以及中继的选择方案简单, 并且能够满足中断概率的最小化, 但是中继节点最好是随机产生的, 不一定都分布在资源和目的节点的连线上。

Ramesh Annavajjala 等^[7]给出了在协作系统中的一种多固定中继选择及功率分配算法, 该方案随机选取 M 个中继节点, 并假设各数据流的最大功率为 P_{\max} , 源节点的功率为 $P_{ii} = \delta_0 P_{\max}$, 剩下的功率在各中继节点中平均分配, $P_{ji} = (1 - \delta_0) P_{\max} / M$, 采用与单中继相类似的算法可以获得理想的功率分配, 由此可以相应的求出中断概率, 采用该方案的确可以获得中断概率的极小值, 但是 δ_0 的计算过程相当复杂, 与 Zhou Kenan 等^[6]按照距离分配中继功率不同, 他们采用平均分配的方法, 但是由于各中继节点的位置不同, 其相应的平均信道增益也不尽相同, 而该方案给条件好的节点和较差的节点都分配相同的功率, 不能实现系统效率的最大化, 因此并不是最佳的分配方案。Tae Won Ban 等^[8]给出了在 DF 模式下的两种多中继选择策略——固定中继选择和自适应中继选择 (平均信道增益低于门限值的中继不参与协作), 功率采用平均分配的方法且总功率并不受限, 在中断概率不断变化的情况下, 又是一种 DF 协作模式下的中继选择算法, 并且设定了一个平均信道增益的门限值, 可以较好的选择中继参与协作, 与 Ramesh Annavajjala 等^[7] 算法相比, 该方案总是能够使用更少的中继, 达到相同的效果, 系统利用率较高, 但是该算法门限值的选取比较麻烦, 中继选择的过程也太过复杂, 系统消耗较大, 同时也未对功率进行合理的分配。Goldsmith 等^[9]给出了在快速瑞利衰落的信道中的 DF 模式下的两种单协作中继的功率分配策略: 第一种为平均功率分配, 第二种为实际功率分配, 在该方案中功率分配方式简单, 计算量较小, 能够实现较高的传输速度并提升系统容量。Qiulin 等^[10]中给出了 DF 模式下的中继选择和功率分配策略, 其中继选择过程相对比较简单, 只需要对所有节点的平均信道增益值进行比较, 能够较好提升网络使用时间, 减少整体功率消耗。

但是在实际情况中, 特别是对于蜂窝无线系统、无线局域网、无线城域网等各种商用的无线系统, 中继节点基本上处于移动的过程当中的, 而中继节点的移动会大大降低系统的性能, 需要重新选择新的中继并实施切换, 但目前该方面的结果很少。本文探讨协作分集中的移动中继选择算法, 在 AF (放大转发) 协作通信模式下, 给出了基于瞬时信道状态信息的功率分配及中继的动态选择策略。针对单中继情况提出了中继切换方案, 针对多中继情况提出了动态剔除、补充中继的方案, 并导出了分集增益及系统容量的计算公式。通过模拟仿真分析, 该方案能有效的降低中断概率, 提高系统分集增益, 扩大系统容量, 实现良好的整体性能, 从而提高了移动中继协作系统的可靠性。

1 单中继节点的选择策略

当协作系统中的中继移动时, 中继与源节点以及目的节点之间的距离会随之发生变化, 如果继续适用原中继进行通信, 系统的误码率 (BER) 和中断概率会随之逐渐上升, 同时系统容量以及分集增益会逐渐下降, 这就降低了系统的性能; 资源节点、中继节点以及目的节点都有各自的覆盖范围, 一旦中继运动到该范围之外, 通信就会中断, 因此在单个移动中继的协作系统中, 需要考虑协作中继的切换问题。这里主要研究在中继移动时, 考虑根据瞬时信道增益的变化来制定相应的中继选择策略及功率分配方式, 以获得理想的中断概率。

中继移动时, 其周围环境随时变化, 由于无线电信号的反射、绕射以及散射, 会带来信号的大尺度损耗, 而无线信道的多径传播将会带来小尺度衰落, 具体表现为: ①经过短距和短时传播后信号强度的急剧变化; ②存在时变的多普勒频移, 引发随机的频率调制; ③多径传播时延会引起时间弥散。这就严重影响了无线系统的可靠性。本文假设资源节点 (源节点) 和目的节点都处于静止状态, 只有协作中继处于运动中, 因此衰落只与空间路径有关^[4]。

假设在一个无线通信网络中, 发送信息的资源节点集合为 $S = \{1, 2, \dots, M\}$, 对于资源节点 $i (i \in S)$ 而言存在着多个中继节点 j , 如果中继数量较大, 不可能对所有的中继信息都进行比较, 因此源节点只从中继节点中随机选取几个进行比较, 采用条件最好的节点进行协作, 然后向中继节点及目的节点发送信息。

假设系统的总带宽为 B ; 带宽归一化的传输速率为 R ; 互信息量同样用带宽归一化; 假设节点之间的信道为相互独立的频率非选择性信道, 信道慢衰落, 某一时刻源点的发送功率为 P_s , 源点对目的节点的发送功率为 P_1^j , 源点对中继的发送功率为 P_2^j , 中继节点 j 的发送功率为 P_3^j ; 源点与目的节点之间的瞬时信道增益为 0 均值循环对称复高斯随机变量, 其方差为 Ω_1 , 源点与中继 j 之间信道增益的方差为 Ω_2^j , 中继 j 与目的节点的信道增益的方差为 Ω_3^j ; 假设基带信号为独立的高斯分布的随机变量, 均值为 0 , 方差等于各自的瞬时发射功率, 因此源点对目的节点和中继 j 的瞬时信噪比分别为 r_1 、 r_2 , $\bar{r}_1 = E[r_1] = P_s \Omega_1 / \sigma^2$ 和 $\bar{r}_2 = E[r_2] = P_s \Omega_2^j / \sigma^2$, 中继对目的节点的瞬时信噪比为 r_3 , $\bar{r}_3 = E[r_3] = P_3^j \Omega_3^j / \sigma^2$, 其中 σ^2 加性高斯白噪声的功率, $\sigma^2 = NB$; N 为单边谱密度, SNR 为系统信噪比。

在 AF 模式下使用单个协作中继策略, 目的节点采用最大比合并的方法对接收信号进行处理, 系统的瞬时互信息量为^[7]:

$$I_i = \log_2[1 + r_1 + r_2^j r_3^j / (1 + r_2^j + r_3^j)] / 2 \quad (1)$$

数据流的中断概率为^[7]

$$P_{out} = P(I_i < R) = P[r_1 + r_2^j r_3^j / (1 + r_2^j + r_3^j) < 4^R - 1] \quad (2)$$

在大信噪比的情况下, 上式可以近似表达成:

$$P_{out} = c_{AF,1} (1/P_s) [(1/P_s + (\alpha_{ij}/P_3^j))] \quad (3)$$

其中 $c_{AF,1} = (4^R - 1)^2 \sigma^4 / (2\Omega_1 \Omega_2^j)$, $\alpha_{ij} = \Omega_2^j / \Omega_3^j$ 。

观察 (3) 可以看出: 中断概率的问题就转化为了功率分配的问题, 而瞬时信道增益也是一个变量, 因此要实现中断概率的最小化, 就是在总功率一定的情况下, 利用瞬时信道增益来选择适当的功率分配准则。

$$P_s + P_2^j = P_{\max} \tag{4}$$

$$P_s = \delta_0 P_{\max} \tag{5}$$

其中 $\delta_0 = (\Omega_1 + \Omega_2^j)/(\Omega_1 + \Omega_2^j + \Omega_3^j)$ 。

对于单个移动中继的情况而言，移动中继的切换实际上就是保证协作的中继节点在源点和目的节点的覆盖半径以内，使通信能够顺利进行。于是设定切换的判决门限，该门限值的选取取决于协作中继节点的位置、移动的速度（也就是信号的衰减速率）以及驻留时间来考虑；切换时间由话务量的大小和寻找中继所需的平均时间决定。

因此，单中继系统的中继选择步骤如下。

初始状态时随机选取几个中继为候选中继，比较各自的中断概率，选取中断概率最小的中继作为协作中继；

当中继的 Ω_2^j 或者 Ω_3^j 低于寻找门限时，源点就开始寻找新的中继，其过程如（1）所述，一旦中继的 Ω_2^j 或者 Ω_3^j 低于切换门限时，系统就进行中继切换。

中继的选取就是循环进行步骤（1）、（2）维持正常的通信。需要特别说明的是：如果系统对于中断概率有具体要求时，步骤（1）在第 1 次选取后没有找到合适的中继，系统可以去掉这些中继，再次重复上述步骤，直到找到适当的中继为止。

2 多中继节点的选择策略

假设系统工作于频分多址（FDMA）环境下，共有 m 个用户参与协作，所有用户均为半双工模式，且可以同时 m 个频率上收发信息。其传输过程如图 1 所示，包括：①源点向所有预选的中继广播信息，②被选择的中继转发信息。

第一步		时 间 轴			
		第二步			
频 率 轴	1 广播	2 转发 1	3 转发 1	m 转发 1
	2 广播	1 转发 2	3 转发 2	m 转发 2

	M 广播	1 转发 m	2 转发 m	m-1 转发 m

图 1 多中继协作示意图

从图中可以看出,在整个协作周期中,为了保证各个协作用户在正交的子信道上传输数据,每个用户台使用时频交错的子信道分别发送自己和其他用户的信息,所有关于用户的信息都只在该用户的多址信道上传输。下面分析中所用符号与第 1 节相同。

2.1 多中继节点的选择策略

在 AF 模式下使用多个协作中继策略,目的节点采用最大比合并的方法对接收信号进行处理,假设存在 M 个协作中继,其集合为 Φ_l , 系统的瞬时互信息量为^[7]:

$$I_i = \log_2[1 + r_1 + \sum_{j \in \Phi_i} r_2^j r_3^j / (1 + r_2^j + r_3^j)] / (M + 1) \quad (6)$$

数据流的中断概率为^[7]:

$$P_{out} = P(I_i < R) = P[r_1 + \sum_{j \in \Phi_i} r_2^j r_3^j / (1 + r_2^j + r_3^j) < 2^{(M+1)R} - 1] \quad (7)$$

在大信噪比的情况下, 上式可以近似表达成:

$$P_{out} = c_{AF,M} (1/P_s) \prod_{j \in \Phi_i} [(1/P_s + (\alpha_{ij}/P_3^j))] \quad (8)$$

其中 $c_{AF,M} = [(2^{(M+1)R} - 1)^2 \sigma^2]^{M+1} / [(M+1)! \Omega_1 (\prod_{j \in \Phi_i} \Omega_2^j)]$, $\alpha_{ij} = \Omega_2^j / \Omega_3^j$ 。

同样, 多中继中断概率的问题也转化为了功率分配的问题。

$$P_s + P_r = P_{\max} \quad (9)$$

$$P_s = \delta_0 P_{\max} \quad (10)$$

$$\text{其中} \quad \delta_0 = (\Omega_1 + \sum_{j \in \Phi_i} \Omega_2^j) / (\Omega_1 + \sum_{j \in \Phi_i} \Omega_2^j + \sum_{j \in \Phi_i} \Omega_3^j) \quad (11)$$

$$P_3^j = (\Omega_3^j / \sum_{j \in \Phi_i} \Omega_3^j) P_r \quad (12)$$

在多个移动中继的协作系统中, 由于性能较差的中继也有可能被选中 (以后简称坏点), 系统的中断概率会随着坏点数目的增加而变大, 因此必须设计一个中继选择方案来排除不良中继。假设数据流在 M 个协作中继时其中断概率为 $P_{out}(M)$, 中继节点的集合为 Φ_i ; 在 $M-1$ 个中继进行协作时, 其中断概率为 $P_{out}(M-1)$, 中继节点的集合为 Φ_i' , 要使去掉一个中继节点后的中断概率降低, 也就是要保证

$$P_{out}(M) / P_{out}(M-1) > 1 \quad (13)$$

带入中断概率公式可以得到:

$$\frac{P_{out}(M)}{P_{out}(M-1)} = \frac{(2^{(M+1)R} - 1)^{M+1}}{(2^{MR} - 1)^M (K+1)} \frac{(1/P_s + \alpha_{il}/P_3^l) \sigma^2}{\Omega_2^l} \frac{P_s'}{P_s} \prod_{j \in \Phi_i'} \frac{P_s'^{-1} + \alpha_{ij} P_3'^{-1}}{P_s' + \alpha_{ij} P_3'^{-1}} \quad (14)$$

分别令

$$A = \frac{(2^{(K+1)R} - 1)^{K+1}}{(2^{KR} - 1)^K (K+1)}$$

$$B = \frac{(1/P_s + \alpha_{il}/P_3^l) \sigma^2}{\Omega_2^l}$$

$$C = \frac{P_s'}{P_s} \prod_{j \in \Phi_i'} \frac{P_s'^{-1} + \alpha_{ij} P_3'^{-1}}{P_s' + \alpha_{ij} P_3'^{-1}}$$

由于去掉中继后, 其功率按比例分配给了剩下的所有节点, 因此 $P_s' > P_s, P_3' > P_3$, 所以 $C > 1$,

$$B = \frac{(1/P_s + \alpha_{il}/P_3^l) \sigma^2}{\Omega_2^l} = \left[\frac{1}{\Omega_2^l \delta_0} + \frac{\sum_{j \in \Phi_i} \Omega_3^j}{(1 - \delta_0)(\Omega_3^l)^2} \right] \frac{\sigma^2}{p_{\max}}$$

$$Q \ a + b \geq 2\sqrt{ab}, \quad \therefore B \geq \frac{2\sigma^2}{\Omega_3^l p_{\max}} \sqrt{\sum_{j \in \Phi_i} \Omega_3^j / [\delta_0 (1 - \delta_0) \Omega_2^l]}$$

$$\text{又 } Q_{ab} \leq \frac{(a+b)^2}{4}, \therefore B \geq \frac{4\sigma^2}{p_{\max} \Omega_3^l} \sqrt{\sum_{j \in \Phi_i} \Omega_3^j / \Omega_2^l}$$

$$\text{因此要保证 } AB > 1, \text{ 只要 } A \frac{4\sigma^2}{p_{\max} \Omega_3^l} \sqrt{\sum_{j \in \Phi_i} \Omega_3^j / \Omega_2^l} > 1$$

$$\text{令 } SNR = \frac{p_{\max}}{\sigma^2}, \text{ 可以化简上式为 } \frac{\Omega_2^l (\Omega_3^l)^2}{\sum_{j \in \Phi_i} \Omega_3^j} \leq 16A^2 / SNR^2$$

所以令门限值, 只要当中继节点 l 的参考值 $\frac{\Omega_2^l (\Omega_3^l)^2}{\sum_{j \in \Phi_i} \Omega_3^j}$ 小于门限的时候, 就出去该节点以

降低中断概率。

因此, 多中继系统的中继选择步骤如下。

(1) 初始状态时随机选取 n 个中继为候选中继, 使所有的候选中继都参与协作;

(2) 计算各中继的参考值 $g = \frac{\Omega_2^l (\Omega_3^l)^2}{\sum_{j \in \Phi_i} \Omega_3^j}$ 的大小, 并按从小到大排序, 其最小值设为 g_{\min} ;

(3) 如果 $g_{\min} < G$, 则去掉该中继, 更新 Φ_i , 否则转步骤 (5);

(4) $M = M - 1$, 更新门限值 G , 如果 $M = 1$, 转步骤 (5), 否则执行步骤 (2);

(5) 结束。

上述过程为某一时刻段除去坏点的过程, 整个数据流的传输过程为:

(1) 在 t_1 时刻随机选取 n 个中继为候选中继, 使所有的中继都参与协作;

(2) 执行上述除去坏点的过程, 此时剩下 m 个中继节点进行通信;

(3) 每隔一段时间 t 更新一次使用中的中继数据, 并完成一次坏点排除过程;

(4) 当中继节点个数为 1 或者中断概率低于相应的通信参考门限时, 立刻随机加入 $n - m$ 个中继节点, 并执行除去坏点的过程, 如此反复直到通信结束。

选取一个合适的时间间隔 t , 使得在该时间段内信道各项参数变化不是很大, 所以其间的各项参数取 t_1 时刻的值, 直到下一个时刻 t_2 到来时再更新所有数据。同样根据不同的地域特征选择相应的 n 值。

由于引入了参考值 g 和门限值 G 的比较, 使得系统不用把每个节点的中断概率都计算出来, 否则系统每计算一次就更新所有的数据, 然后再重新计算, 那样大大增加了系统的负担, 降低了运行效率。

2.2 中断概率受限时中继的选择

如果系统对中断概率有特殊要求时, 随机选取的中继节点可能并不满足通信的需求, 这样长时间的寻找中继会带来较大的时延甚至掉线, 因此对于中继受限时需要首先对中继节点的信息进行预处理, 选出一定数量的合适的中继以供选择。

单协作中继时, 根据式 (3) 系统的中断概率为

$$P_{out} = c_{AF,1} (1/P_s) [(1/P_s + (\alpha_{ij}/P_3^j))]$$

将各项参数带入可以得到

$$P_{out} = \frac{(4^R - 1)^2 \sigma^4}{2\Omega_1 \Omega_2} \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2} \frac{1}{P_{max}} \left(\frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2} \frac{1}{P_{max}} + \frac{\Omega_2}{\Omega_3} \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_3} \frac{1}{P_{max}} \right)$$

整理后得到

$$P_{out} = \frac{(4^R - 1)^2 \sigma^4}{2\Omega_1 P_{max}^2} \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2} \left(\frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2} + \frac{\Omega_2}{\Omega_3} \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_3} \right) \frac{1}{\Omega_2}$$

分别令

$$A = \frac{(4^R - 1)^2 \sigma^4}{2\Omega_1 P_{max}^2}$$

$$B = \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2}$$

$$C = \left(\frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2} + \frac{\Omega_2}{\Omega_3} \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_3} \right) \frac{1}{\Omega_2}$$

其中 A 为常数,

$$C = \left(\frac{1}{\Omega_1 + \Omega_2} + \frac{\Omega_2}{\Omega_3^2} \right) \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_2}$$

$$Q \ a + b \geq 2\sqrt{ab}, \quad \therefore C \geq 2\sqrt{\frac{\Omega_2}{\Omega_1 + \Omega_2} \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_2 \Omega_3}}$$

$$\text{整理得 } C \geq \frac{2}{\Omega_3} \sqrt{\frac{\Omega_1 + \Omega_2}{\Omega_2}} + \frac{2}{\Omega_2} \sqrt{\frac{\Omega_2}{\Omega_1 + \Omega_2}}$$

$$Q \ a + b \geq 2\sqrt{ab}, \quad \therefore C \geq \frac{4}{\sqrt{\Omega_2 \Omega_3}}$$

$$\therefore BC \geq \frac{\Omega_1 + \Omega_2 + \Omega_3}{\Omega_1 + \Omega_2} \frac{4}{\sqrt{\Omega_2 \Omega_3}}$$

整理得

$$\therefore BC \geq \frac{4}{\sqrt{\Omega_2 \Omega_3}} + \frac{\Omega_3}{\Omega_1 + \Omega_2} \frac{4}{\sqrt{\Omega_2 \Omega_3}}$$

$$Q \ a + b \geq 2\sqrt{ab}, \quad \therefore BC \geq \frac{8}{\sqrt{(\Omega_1 + \Omega_2)\Omega_2}}$$

假设中断概率最大值为 P , 利用不等式 $P_{out} \geq P$ 可以求取相应参数的取值范围,即

$$P_{out} = ABC \geq \frac{8A}{\sqrt{(\Omega_1 + \Omega_2)\Omega_2}} \geq P$$

也就是需要

$$(\Omega_1 + \Omega_2)\Omega_2 \leq \frac{64A^2}{P^2} \quad (15)$$

所以令门限值 $H = \frac{64A^2}{P^2}$, 从 (15) 式不难看出当中继节点的参考值 $(\Omega_1 + \Omega_2)\Omega_2$ 小于门

限值的时候, 系统的中断概率大于受限中断概率值, 不能满足通信需求, 而反过来取参考值 $(\Omega_1 + \Omega_2)\Omega_2$ 大于门限值 H , 大部分节点都能满足通信需求, 由于是通过不等式放缩得到的参考门限, 所以得到的中继节点不一定完全满足 $P_{out} \leq P$, 但是前文中假设的是单中继通信的情

况，而采用多中继通信时，中继条件的要求不如单中继那么苛刻，这些在单中继受限时不能使用的中继可以参与多协作中继通信，同样可以满足中断概率的需求，因此中断概率受限时中继的选择过程如下。

(1) 在 t_1 时刻根据受限中断概率值计算出门限值 H ，选出 n 个满足该门限的中继；

(2) 从 n 个中继中随机选择两个中继并计算出其中断概率，如果满足条件，则执行步骤(3)，如果不满足条件，就增加协作中继个数，同时执行除坏点的过程，直到获得的中断概率满足要求为止（为了避免中继移动后性能恶化较快而不停地进行中继重选，可以使获得的中断概率值远小于受限值，相应比例幅度根据具体的移动环境而定）；

(3) 每隔一段时间 t 更新一次在用的及备用的中继数据，完成一次坏点排除过程，查看现有系统的中断概率是否逼近受限中断概率的临界值 KP ，其中 $K \in (0,1)$ （其大小视具体的移动环境而定）；

(4) 当中继节点个数为 1 或者中断概率高于 KP 时，立刻执行步骤(2)如此反复直到通信结束。

3 分集增益及系统容量的计算

这里针对多中继情况给出分集增益以及系统容量的计算，单中继为其特例 ($M=1$)。虽然文中没有介绍编译码的方案，但是纠错编码的分集增益 G_c 可以通过下式求得近似值^[7]。

$$G_c = \frac{\delta_0}{2^{(M+1)R} - 1} [(M+1)\Omega_1 \prod_{j \in \Phi_i} \frac{\Omega_2^j \delta_j}{\delta_j + \delta_0 \alpha_{ij}}]^{\frac{1}{M+1}} \quad (16)$$

$$P_{out} \approx (G_c \cdot SNR)^{-G_d} \quad G_d \approx -\frac{\lg P_{out}}{\lg(G_c \cdot SNR)} \quad (17)$$

其中 G_d 为分集增益， δ_j 为分配给第 j 个中继的功率系数。

当采用固定中继的时候，编码增益 G_c 为一个常数，因此分集增益 G_d 只是信噪比的函数。但是，当中继移动时，参与协作的中继数量在不断变化，这就会引起各条链路的平均信噪比的变化，功率分配系数也随之改变，由此可以看出，移动中继下的编码增益及分集增益都是多个变量的函数，并在不停的变化中。

根据(6)式，系统的容量为：

$$C = \max(I) = \max \log_2 [1 + r_1 + \sum_{j \in \Phi_i} r_2^j r_3^j / (1 + r_2^j + r_3^j)] \quad (18)$$

由于在一个固定时刻，各点瞬时信噪比都是固定的，因此整个系统的容量为：

$$C = \log_2 [1 + r_1 + \sum_{j \in \Phi_i} r_2^j r_3^j / (1 + r_2^j + r_3^j)] \quad (19)$$

将前文所述的瞬时信噪比的计算公式带入(18)式，化简整理可以得到带宽归一化系统容量的表达式为：

$$C = \log_2 [1 + \delta_0 \cdot SNR \cdot \Omega_1 + \sum_{j \in \Phi_i} \frac{\delta_0 (1 - \delta_0) \Omega_2^j (\Omega_3^j)^2 SNR^2 / \sum_{j \in \Phi_i} \Omega_3^j}{1 + \delta_0 \Omega_2^j SNR + (1 - \delta_0) (\Omega_3^j)^2 SNR^2 / \sum_{j \in \Phi_i} \Omega_3^j}] \quad (20)$$

同样从上式可以看出，固定中继的信道容量是一个常数，它不会随着时间的变化而变化；

而在中继移动时，当参与协作的信道数目和中继的相对位置发生变化时，信道容量也会随之改变。

4 仿真实验结果

假设仿真中的信道环境为瑞利平坦衰落信道，任意两节点之间的瞬时信道增益服从均值为 Ω 的瑞利分布，其大小用公式 $\Omega = k/(L/D)^4$ 进行计算，其中 L 为任意两节点之间的距离， D 为源点与目的节点之间的距离，一旦中继与源点或者目的节点的距离超过了 D 就更换中继，系数 4 为路径损耗系数， k 为增益系数，由两节点的发射和接收天线的增益、高度、工作环境以及工作频率决定，对于固定的源点和固定中继而言， k 保持不变。

4.1 中继未移动时中断概率随信噪比的变化

仿真参数选取如下：

$M = 5$ ， $R=1\text{bit/sec/Hz}$ ，

资源节点同目的节点的距离 $D=50\text{m}$ ，增益系数 $k=1$ ，

每条链路的噪声功率相等，

多中继协作时，随机产生 5 个节点 $ABCDE$ ， $ABCD$ 的运动方向分别是上下左右，节点的移动速度 $v=10\text{m/s}$ ， E 点不动；单中继协作时，任意从 $ABCD$ 中取 1 点作为其中继节点。

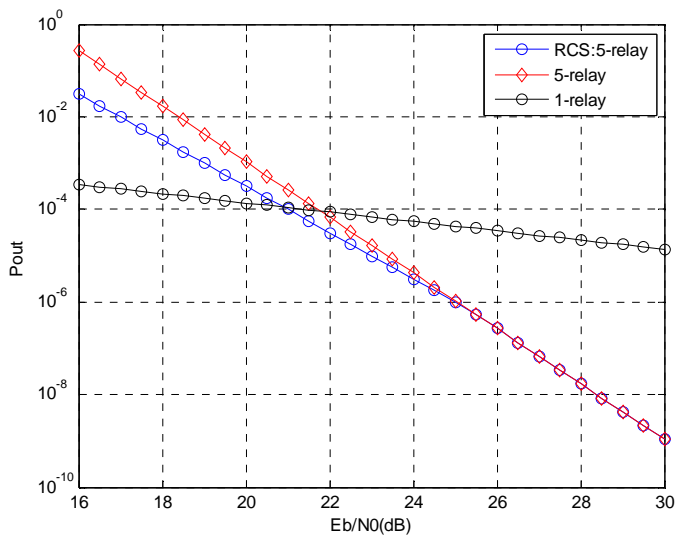


图 2 中继未移动时协作系统中断概率随信噪比的变化图

当中继移动时间 t' 后，对采用中继选择策略(RCS: 5-relay)、固定 5 中继不变(5-relay)以及单中继(1-relay)三种协作策略的性能进行比较。从图 2 来看，随着信噪比的提升，三者的中断概率都有所降低；当系统信噪比较低时，单中继协作系统所受到的干扰远小于多中继协作系统，因此其中断概率自然比多中继的小，而随着信噪比的增加，多中继的优势就逐渐体现了

出来；在低信噪比时，多中继选择策略按照参考值 g 从大到小的顺序参与协作，在任何时刻都不包含坏点，因此其中断概率要低于固定 5 中继不变情况，当信噪比提升到 25db 时，所有的节点都参与协作，两者获得了相同的分集增益，所以中断概率最后都相同。其实采用中继选择策略时，每增加一个中继节点，系统的中断概率都会出现一个跳点，这是由于这个坏点的剔除门限是用不等式求出的，其对于消除坏点有效，但是对于增加中继是有误差的，所以在做仿真的时候，协作中继的数量每变更一次都对前后的中断概率进行了比较，以确定是否增加了该中继能够有效地降低中断概率。但是在实际的应用过程中，在任何一个时间间隔 t 中，系统的 SNR 不会发生如此大的波动，所以在该间隔内不用考虑增加协作中继的情况，因此对实际的应用不会产生影响。

4.2 中继移动时中断概率随时间的变化

仿真参数选取如下：

$M = 4$, $R=1\text{bit/sec/Hz}$,

资源节点同目的节点的距离 $D=50\text{m}$, 增益系数 $k=1$,

每条链路的噪声功率相等,系统总信噪比固定为 25db,

多中继协作时，随机产生 4 个节点 $ABCD$, $ABCD$ 的运动方向分别是上下左右，节点的移动速度 $v=10\text{m/s}$, 单中继协作时，任意从 $ABCD$ 中取 1 点作为其中继节点。每隔 1s 更新一次各项参数，当节点超过覆盖半径时，去掉这些中继节点，随机产生 1 个或者几个节点补充，使系统中一直保持 4 个中继参与协作。

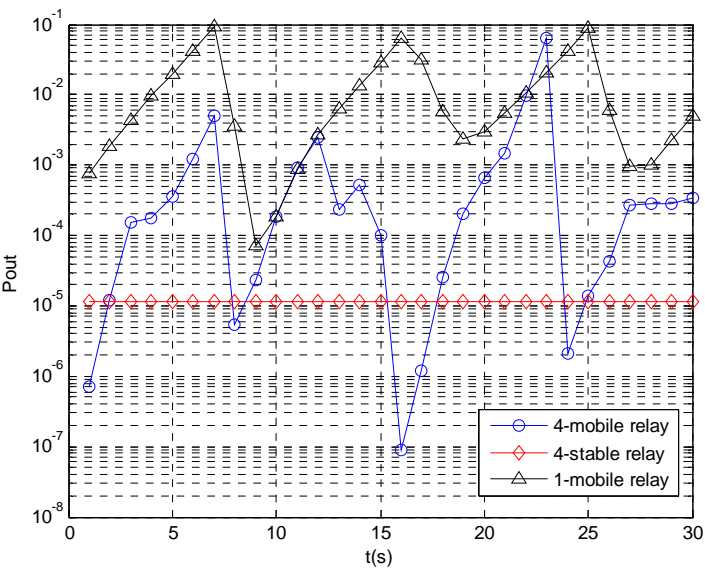


图 3 移动中继与固定中继在 SNR 为 25db 时中断概率的比较

图 3 是移动中继与固定中继在 SNR 为 25db 时中断概率的比较图。由于是在大信噪比的情况下，4 中继的中断概率初始值为到 10^{-5} ，远小于单中继的 10^{-3} ；当中继移动 8s 后，碰巧单中继和多中继系统中都有中继节点不能继续工作，于是第 9s 时，单中继进行了中继切换，中

断概率从 0.093 降低到 0.0012，多中继的中断概率值从 0.005 降低到 5.3×10^{-6} ，进行中继切换和重选后系统性能提升明显，能够维持一段时间直到下一次中继切换或者重选；同时随着移动中继相对位置的变化，中断概率值也在不停地跳变中，但是其大小总是在固定中继的中断概率附近跳跃。

4.3 移动中继中断概率受限时中断概率随时间的变化

要使图 3 中的 4 移动中继的中断概率限制在 10^{-2} 以下，系统需要首先对中继数据进行处理，然后再从备选的中继中选取合适的中继进行通信。如图 4 所示：未受限的移动中继在第 7s 时才开始准备更换中继，而在中断概率受限的时候，为了防止在下一秒内中继移动后系统性能超过受限值从而产生掉线的情况，于是系统设定中断概率 10^{-3} 为一个保护值，一旦超过该值就考虑更换中继，避免了性能恶化造成掉线，于是在 6s 时就准备切换中继，在第 7s 时系统性能得到明显改善，在一定的时间段内系统会相对稳定，如果周围的中继条件都不理想，切换后还是不能满足通信需求，那么就增加中继个数以维持系统的稳定性。在整个时间段内，系统中断概率总是小于 $10^{-2.5}$ ，能够满足通信的需求。

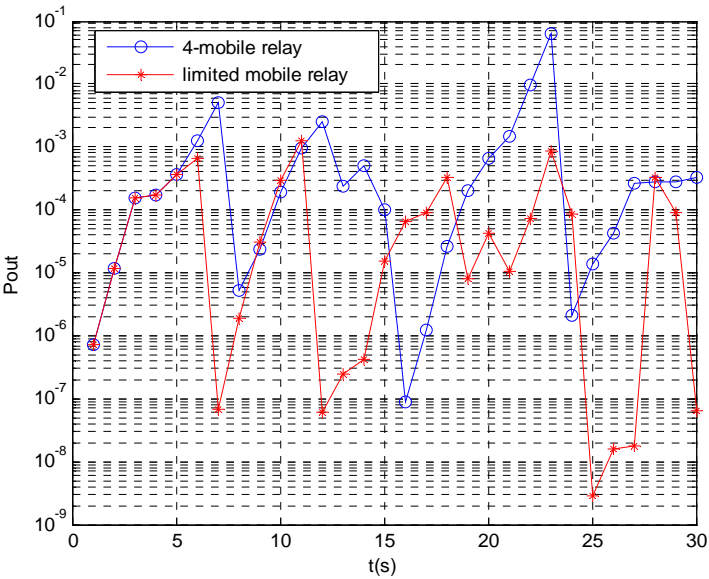


图 4 移动中继中断概率受限时中断概率随时间的变化

4.4 分集增益的比较

图 5 为 1 个移动中继的协作系统(1-mobile relay)与随机选择的 4 个移动中继的协作系统(4-mobile relay)的分集增益比较。从图中可以明显看出，多中继协作系统的增益明显高于单中继的协作系统。

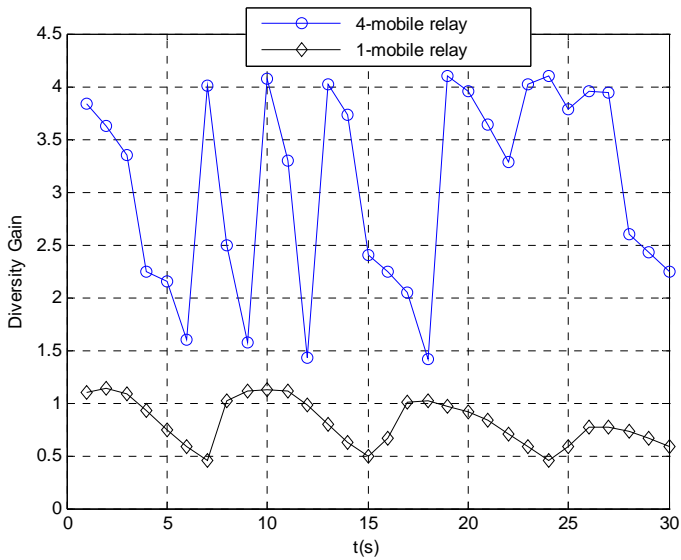


图 5 分集增益的比较

4.5 系统容量的比较

图 6 为各类中继在不同情况下的系统总容量比较。由于多中继系统会根据具体情况剔除或者补充中继节点，因此其容量值总是在固定中继容量值附近徘徊，并且数值上一般都小于固定中继。单移动中继较多中继少了三个信道，中继个数决定了系统容量，因此单中继的系统容量较多中继要小很多。

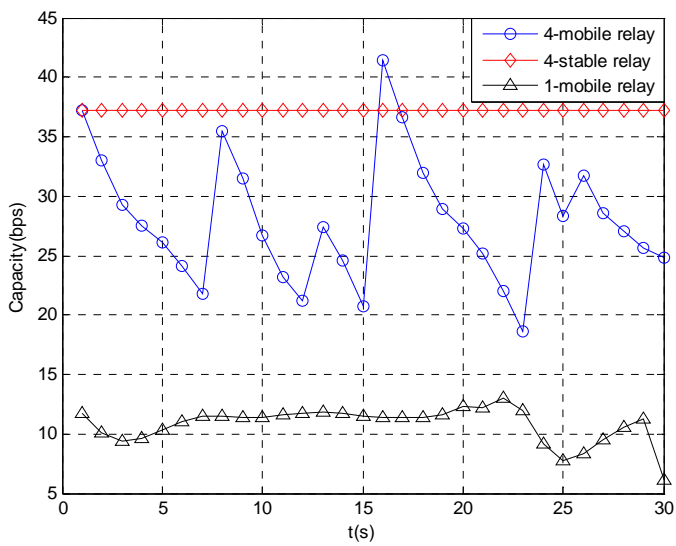


图 6 系统容量的比较

5 结论

本文针对放大转发协议,给出了基于瞬时信道状态信息的功率分配及中继的动态选择策略。针对单中继情况提出了中继切换方案,针对多中继情况提出了动态剔除、补充中继的方案,并导出了分集增益及系统容量的计算公式,实现良好的整体性能。下一步工作将分析其他协作协议的性能和研究多跳的情况。

参 考 文 献

- [1] Sendonaris A, Erkip E, AAZhang B. User cooperation diversity; part I: system description [J]. IEEE Transaction on Communications, 2003,51(11); pp1927-1938.
- [2] Sendonaris A, Erkip E, AAZhang B. User cooperation diversity; part II: implementation aspects and performance analysis [J]. IEEE Transaction on Communications, 2003, 51(11); pp1939-1948.
- [3] Laneman, J. N., Tse, D. N. C., and Wornell, G.W. Cooperative diversity in wireless networks: Efficient protocols and outage behavior[J]. IEEE Trans. Inform. Theory, 2004; pp:3062–3080.
- [4] Hunter, T. E., Nosratinia, A.: Diversity through coded cooperation. Wireless Communications[J], IEEE Transactions on Volume 5, Issue 2, Feb. 2006 ; pp:283—289.
- [5] 雷维嘉、谢显中、李广军, 一种基于 LDPC 编码的协作通信方式[J], 电子学报, 2007; 35(4) pp:712-715.
- [6] Zhou Kenan, Tat Ming Lok, A Relay Selection Scheme under Optimal Power Allocation[C], IEEE ICCS 2008; pp:1609-1613.
- [7] Annavajjala R, Cosman P C, Milstein L B. Statistical channel knowledge-based optimum power allocation for relaying protocols in the high SNR regime [J]. IEEE Journal on Selected Areas in Communications, 2007, 25(2); pp292-305.
- [8] Tae Won Ban, Bang Chu1 Jung, Dan Keun Sung, Wan Choi, Performance analysis of two relay selection schemes for cooperative diversity[C], The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007.
- [9] Chris T.K. Ng, Andrea J. Goldsmith, The Impact of CSI and Power Allocation on Relay Channel Capacity and Cooperation Strategies[J], IEEE Transactions on Wireless Communications, VOL. 7,NO.12, Dec.2008; pp:5380-5389.
- [10] Qiulin, Zhang Yong, Daichao, Song Mei, Wang Danzhi, Cross-layer Design for Relay Selection and Power Allocation Strategies in Cooperative networks[C], The Seventh Annual Communication Networks and Services Research Conference; 2009, pp:232-237.
- [11] 廖晓光. 非对称中继协作的中断容量与资源分配[D]. 中国科学技术大学. 2008.5.1
- [12] Yilong Zou. Baoyu Zheng. Jia Zhu. Out analysis of opportunistic cooperation over Rayleigh fading channels. [J]. IEEE Transaction on Wireless Communications, VOL. 8, NO.6, 2009.6.; pp3077-3085.
- [13] Alejandro R. Xiaodong Cai. Georgios B. Giannakis. Symbol error probabilities for general cooperative links. [J]. IEEE Transactions on Wireless Communications, VOL. 4, NO.3, 2005.5; pp1264-1273.
- [14] 谢显中、雷维嘉, 移动通信中的空时信号处理[M], 北京: 电子工业出版社, 2008.
- [15] Kramer G, Maric I. and Yates R.D., Cooperation Communications[J], Foundations and trends in Networking, Vol.1, No.3-4, 2006, 271-425.

作者介绍

张 鑫(1982-), 男, 四川人, 硕士研究生, 主要研究方向为移动通信。E-mail: zhangxin7722@gmail.com

谢显中(1965-), 男, 四川人, 博士, 教授。主要研究方向为移动通信技术、通信信号处理。E-mail: xiexzh@cqupt.edu.cn

雷维嘉(1969-), 男, 云南元谋人, 副教授。主要研究方向为无线通信技术。E-mail: leiwj@cqupt.edu.cn

第 4 部分

网络理论与技术

基于无线传感器网络的危险货物运输系统

陈 晨¹ 裴庆祺¹ 庞辽军¹ 张素兵² 范科峰²

(1.西安电子科技大学综合业务网国家重点实验室,西安 710071;

2.中国电子技术标准化研究所,北京 100007)

摘 要: 本文分析了危险货物运输系统所面临的安全威胁,并提出了采用无线传感器网络进行监控的方法。随后给出了具体的实现架构、原型系统、安全模型和控制中心设计,通过详细的论述验证了该方案的可行性与有效性。

关键词: 危险货物; 运输系统; 无线传感器网络

Transportation System For Hazardous Goods Based On WSN

CHEN Chen¹ PEI Qing-qi¹ PANG Liao-jun¹

ZHANG Su-bing² FAN Ke-feng²

(1.National Key Lab. of Integrated Service Networks, Xidian Univ., Xian 710071;

2.China Electronics Standardization Institute, Beijing 100007)

Abstract: Based on the analysis to the security threats in hazardous goods transportation, the wireless sensor networks have been introduced for monitoring and supervision. Then, we give the detailed implementation infrastructure; prototype system, security model and the design of control center. Finally, through thorough argumentation, the feasibility and validity of our solution have been verified.

Keywords: hazardous goods; transport systems; wireless sensor networks

引言

“危险货物”定义为具有爆炸、易燃、毒害、感染、腐蚀、放射性等特性,在运输、储存、生产、经营、使用和处置中,容易造成人身伤亡、财产毁损和环境污染而需要特别防护和监测的物质和物品。因此,如何实时监测危险货物在途安全状态将是一个具有重大意义和挑战性的课题。

针对危险货物运输问题除了靠严格的规章制度来进行预先防范以外,普遍采取的另一措施是在危险品运输车辆上安装行车记录仪及 GPS 卫星定位系统来进行车辆的全程同步记录监

控。但这种监控方式并不能做到事故的预先避免和及时报警，并且缺少控制中心和在途车辆之间的控制交互，即控制中心并不能起到实际上的“控制”作用。因此，利用先进的传感器网络技术和主动控制技术来实时保证危险货物在途运输的安全是非常有必要的。

1 危险货物运输研究重点

1.1 无线传感器网络监测

由于危险货物往往具有密闭性，并且摆放位置较为严格，因此使用有线连接显然不能满足要求。而无线传感器网络强大的数据传输与监测能力正好能满足这一需求。与传统的环境监测系统相比，使用无线传感器网络进行环境监测具有三个优点：①传感器节点的体积很小，整个网络只要一次布设，因此传感器网络对被监测环境的影响很小；②传感器网络节点数量多，分布密集，传感器网络具有数据采集量大、精度高的特点；③无线传感器本身具有一定的计算能力和数据存储能力，可以实现较为复杂的监控，传感器节点还具有无线通信能力，可以使得节点协同监控，并且不受布线影响。

1.2 危险货物监测传感器

传感器是一种检测装置，能感受到被测量的信息，并能将检测感受到的信息，按一定规律变换成为电信号或其他所需形式的信息输出，以满足信息的传输、处理、存储、显示、记录和控制等要求。目前较为常用的危险环境监测传感器有爆炸传感器[1-4]、辐射传感器[5]、温度湿度传感器[6]、加速度传感器[7]、压力传感器[8]、振动传感器[9]、气敏传感器[10]等等。而这些传感器目前都有较为成熟的产品在市场上流通。

2 监测系统框架研究

2.1 总体框架

危险货物在运输过程中，由于其体积，数量，运输方式等的不同，会使监控其状态的带有多种功能传感器的无线传感器网络节点呈现出不同的拓扑分布。比如由集装箱运输的分散包装的易燃易爆物，由于其中任意一个货物发生事故都将造成所有货物的燃烧或者爆炸，因此需要分层次分散布置大量传感器节点，而如果采用有线网络布线方式连接各个传感器显然是不现实并且不可靠的。因此，针对不同的运输任务采用不同的拓扑分布和数据采集传输方案显然是有必要的。

此外，考虑到传感器节点低功耗、有限计算处理能力、通信速率较低等特点，在途的多功能显示报警终端必须采用其他设备进行实现。该设备需要有较强的处理计算能力，并可以和传感器网络 sink（汇聚节点）节点进行通信，并且具有接入公共基础通信设施的无线通信模块。此外，该节点还需要相应控制中心的主动控制命令，控制命令执行设备进行相应的动作。因此，拟采用的危险货物在途安全状态监测系统体系结构如图 1 所示。

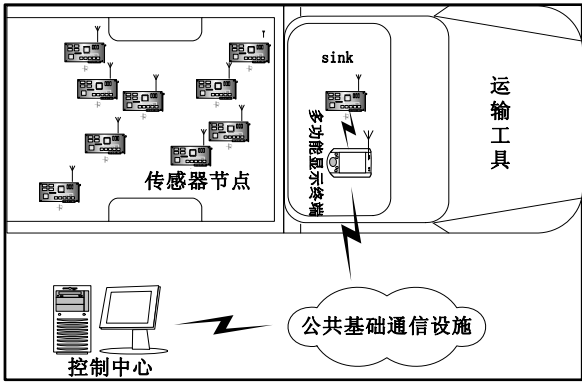


图 1 系统总体结构图

其中传感器节点负责周期性的采集危险货物的安全状态信息并将其传输给 sink 节点, sink 节点将收到的数据进行融合后发往多功能显示终端。多功能显示终端具有较强的存储和计算能力, 其通过预先设定的报警规则或阈值决定是否报警, 并将处理过的数据封装成帧通过公共基础通信设施发送给控制中心。控制中心将收到的带有货物 ID 的状态信息存储在数据库中, 并依据当前和历史的 状态数据进行预测, 并根据预测结果决定是否需要主动干预。如果需要干预, 则将干预指令传给多功能显示终端, 并在在途人员认可的情况下, 执行干预指令。

2.2 原型系统

2.2.1 无线传感器节点

本文设计的无线传感器节点结构图如图 2 所示。

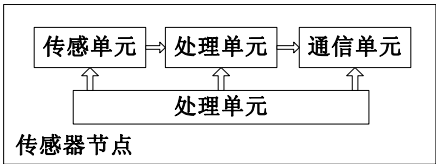


图 2 传感器节点结构图

其中传感单元负责向多功能状态采集传感器提供数据收发接口, 数据经过 A/D 转换后发给处理单元, 处理单元将收到的原始数据按照 IEEE 802.15.4 标准封装成帧, 发送给通信单元。节点的协议结构如图 3 所示。

图 3 中射频收发模块采用专用射频芯片实现, 其具有自动封装/解封 IEEE 802.15.4 物理层帧的能力。该芯片还应该具有信道载波侦听功能, 能够为 MAC 层调度或竞争协议提供参考。数据处理模块采用带有 MCU (微控制器)、片上存储单元、SPI (串行外围设备接口) 或 I2C (集成电路间总线)、串口或 USB (通用串行总线) 等的专用处理芯片或 SOC (片上系统)。数据处理模块和射频收发模块之间采用 SPI 或者 I2C 总线完成数据的双向传输、时钟公用、中断响应等功能, 即图中的层间接口。在数据处理模块中, 首先通过地址映射将 Flash (闪存存储器) 或 EEPROM (电可擦可编程只读存储器) 中的可编程地址进行映射, 供微操作系统使用。然后, 编写具有 ISR (中断服务程序)、任务调度、FIFO (先入先出队列)、定时器等

功能的微型操作系统，在操作系统之上，构建符合 IEEE 802.15.4 的 MAC 层功能，该层负责完成 MPDU（MAC 层协议数据单元）的封装/解封，普通节点和 Coordinator 的初始化，网络安全策略的实施，信标帧的发起和同步，Scan（扫描）和 Association（连接）/Diassociation（断连）等功能。在 MAC 层之上，需要实现具有路由功能的网络层，以支持多跳传输。最顶层是 APPLICATION（应用）层，完成最终数据的融合、进一步处理的工作。

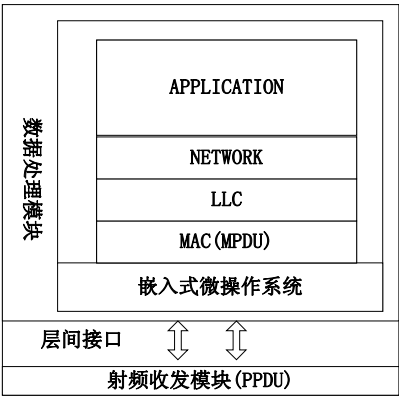


图 3 节点的协议结构图

2.2.2 多功能显示终端

多功能显示终端主要负责给在途人员提供实时数据参考或报警，并将从 Sink 节点收到的数据封装成公用基础通信网络支持的数据帧，发送给控制中心。如果货物产生报警，在途人员可以通过多功能终端上的蓝牙控制器控制车内的辅助设施进行险情控制。多功能显示终端的结构图如图 4 所示。

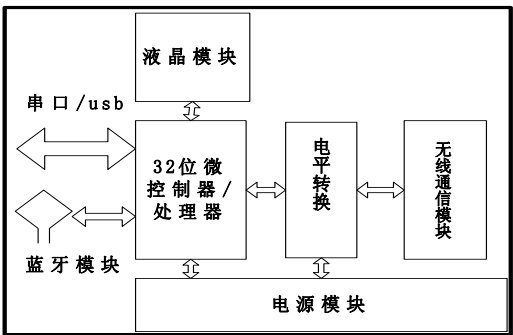


图 4 多功能显示终端

图 4 中核心部分是 32 位的微控制器/处理器模块，它负责将从 sink 收到的数据处理后以图形化的方式显示在液晶屏上，并将数据封装后，经过电平转换模块，传送给无线通信模块。蓝牙模块负责将控制中心发送来的主动控制指令传送给相应的在途安全辅助设施。多功能终端控制辅助设备的结构图如图 5。图中的各种辅助设备，比如灭火器，喷淋器，空调等都配备有蓝牙模块。

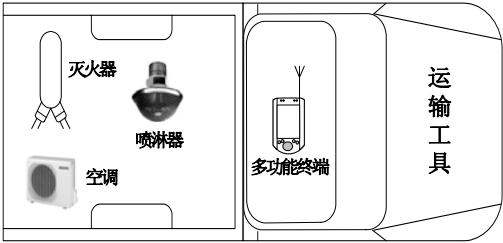


图 5 在途主动控制系统

3 安全模型

3.1 密码算法

传感器网络缺乏网络基础设施，资源受限等特性使得诸多现有的密码算法难以直接应用[11]。目前主要使用是对称密码算法。但是在特定情况下，如访问控制等也使用低开销的非对称密码算法。在对称密码中，消息认证码（MAC）和 hash 被广泛使用，如消息/身份认证通过 MAC 来进行，而不是传统的数字签名方式。广播认证协议 μ TESLA 以及其扩展等基于单向 hash 链的认证协议也能够胜任危险货物的在途运输。

3.2 密钥管理

密钥管理是传感器网络的安全基础。所有节点共享一个主密钥方式不能够满足传感器网络的安全需求。对于危险货物的运输过程，由于传递的信息都是货物的关键状态信息，一旦某个节点的密钥被破解，网络中的数据分组就会被篡改，导致错误的报警信息。因此，采用每个节点与 sink 节点共享一对密钥的方式比较适合我们提出的系统构架。这种方式每个节点需要存储的密钥量小，计算和存储压力集中在配置较高的 sink 节点。该方法计算复杂度也较低，对普通节点资源和计算能力要求不高；引导成功率高；可以支持大规模的传感器网络；sink 能够识别异常节点，并及时地将其排除在网络之外。

3.3 认证技术

由于传感器网络的“一对多”和“多对一”通信模式，广播是能量节约的主要通信方式。因此，传感器网络广播认证具有重要的意义。而针对我们的系统架构，存在着全联通和分簇形式的网络拓扑，使用广播认证既能够有效的鉴别用户的合法性，而且能够节省节点宝贵的能量。

3.4 DoS 攻击防御技术

DoS 攻击就是任何减弱或者消除网络平台期望执行功能的行为。如果攻击者能够发送合法的数据包，就有可能发起 DoS 攻击。因此，这种攻击多来自于网络内部。协作网络监测方法是一种有效的无线传感器网络 DoS 攻击抵御方法。邻居节点之间相互监测，如果在监测时

间 t 内, 没有收到邻居节点的心跳信息, 则产生报警。

4 控制中心设计

图 6 是控制中心系统结构示意图, 无线通信模块将在途运输工具发来的危险货物状态信息接收后转发给分析控制工作站。工作站首先将数据按照预定的格式存入后台数据库, 以供历史数据分析。然后对当前收到的数据进行分析判断, 查看是否有遗漏报警信息的情况。随后, 根据货物 ID 和当前状态信息, 结合其历史数据, 对货物未来状态进行预测, 如果预测趋势显示货物存在安全隐患, 则将预测结果通过无线通信模块发给在途多功能终端。如果有必要, 控制中心也可以发起主动干预指令, 通过多功能终端上的蓝牙控制器来操控在途辅助设备, 进行一些预先处理来减小事故发生的可能性。

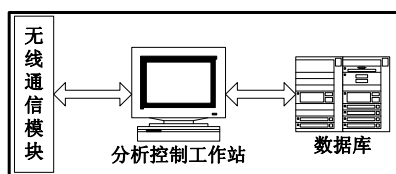


图 6 控制中心系统结构图

5 结束语

本文研究了危险货物运输过程中的关键技术, 包括系统结构、原型系统、安全模型、控制中心等方面, 并给出了采用无线传感器网络进行监测控制的具体实现方法与技术难点。

参 考 文 献

- [1] 孟祥承. 核电子学与探测技术[J], 2003, 3 (4) : 371-379.
- [2] Watson G, Horton W, Staples E. Proceedings of the 1st International Symposium on Explosive Detection Technology[C]. Atlantic City: NJ, 1991: 589-603.
- [3] Cumming C J, Aker C, Fisher M, et al. IEEE Transactions on Geoscience and Remote Sensing[J], 2001, 39 (6) : 1119-1128.
- [4] 新兴市场: 纳米技术传感器[EB/OL]. [2007-11-09]. <http://www.istis.sh.cn/list/list.asp?id=2119>.
- [5] 林友德, 郭亨礼. 传感器及其应用技术[M]. 上海: 上海科学技术文献出版社, 1992: 15-18.
- [6] 郭纯生. 纳米技术-传感器发展的新契机[J]. 传感器技术, 1998, 17 (4) : 5-8.
- [7] 刘迎春, 叶湘滨. 传感器原理设计及应用[M]. 长沙: 国防科技大学出版社, 2002.
- [8] 传感器发展气象万千 [EB/OL]. [2007-11-19]. <http://news.5117.com/news2detail21177.php>.
- [9] 孙利民. 无线传感器网络[M]. 清华大学出版社, 2005.
- [10] 刘威. 气体传感器的研究与发展[J]. 化工纵横, 2000, 14 (9) : 1-11.
- [11] 郑强, 王晓东. 无线传感器网络安全研究[J]. 微计算机信息, 2008, 1 (1) : 116-117.

作者简介: 陈晨 (1977—), 男, 陕西西安人, 西安电子科技大学博士, 主要从事无线近距离通信, 嵌入式系统的研究。

军用ASON融合建网的应用可行性解析

任志宏¹ 谢永强² 赵广松³

(1 重庆通信学院, 重庆, 400035; 2 中国电子设备系统工程公司, 北京, 100041;
3 解放军理工大学, 南京, 210007)

摘要:新型的IP承载网日趋成为面向未来的新一代IP电信网,既要支持语音、视频、文本数据等多媒体业务类型的供求,又要提高网络的服务质量、安全性和可靠性等评价指标的效能。ASON的引入使得传统的光网络具备了动态交换和智能控制特性,促发了光网络从“静态承载”向“动态业务”进行转型,使得兼备各自层级优势融合建网的理念呼之欲出。从“IP Over ASON”的社会可行性、经济可行性以及技术可行性进行了深入分析,根据现实需求,提出了平台建网的应用生存性策略,为军用ASON结构化建设提供参考价值。

关键词: IP Over ASON; 可行性; 生存性策略; 建设

The Applied Feasibility Analysis of Integrated Network Construction for Military ASON

REN Zhi-hong¹ XIE Yong-qiang² ZHAO Guang-song³

(1 Chongqing Communication Institute, Chongqing, 400035;

2 The Electronic Equipment System Engineering Company of China, Beijing, 100041;

3 The University of Science and Technology of PLA, Nanjing, 210007)

Abstract: New bearing network for IP will be next generation IP telecommunication network gradually in the future, which not only supplies multi-media services of sound, video and text, but also improves evaluation index efficiency of QoS, security and reliability. On account of ASON makes the traditional optical network can be switched dynamically and controlled intelligently, the traditional optical network is promoted to transition from “static bearing” to “dynamic service”. Thus, the idea for integrated network combined superiority of different layers is set. The way of social feasibility, economic feasibility and technical feasibility for “IP Over ASON” is analyzed thoroughly. In the light of real demand, the applied survival strategy is advised about integrated network, which is valuable for military ASON construction.

Keywords: IP Over ASON; Feasibility; Survival Strategy; Construction

引言

“信息科学是第一核心生产力”，在这个当今高、精、尖科技文明充盈的大众领域里，人们无时无刻不在享受着信息时代的生活便利。尤其是现代信息通信领域的发展日新月异，信息通信 ALL-IP 和网络融合在未来的网络革新之路中的演进方向已日渐明朗，资源配置的动态化共享，网层结构的扁平化设计，信息服务的安全化保障已成为一种必然而实用的趋势。无论是综合业务本身，还是语音、数据和多媒体的承载方式最终均将实现全 IP 化。对于网络自身的 QoS 认证，可靠性传送，以及生存性规划的基础性指标也将被蕴含新的语义。

随着 IP 技术的成熟，国际标准的修缮，通信网络渐进向全 IP 化网络转型，将形成互操作的，融合的网络结构。这不仅使网络自身实现了平滑演进和高效部署，而且能在有效降低初始预算中的 CAPEX（开销成本）和 OPEX（维护成本）的同时，保护建设方的价值收益。

1 IP网络和光网络的融合转型

1.1 IP网络的发展现状

目前，IP 网络的承载方式主要是 IP over Fiber、IP over DWDM、IP over ATM 以及 IP over SDH 等方式。但是，这几种承载方式均有一个明显的不足，那就是资源利用率很低，保护恢复的方式不够智能多样。甚者，部分承载方式在 IP 网络出现故障时，信息业务的保护恢复仅仅只靠 IP 层自身实现，没有充分启用光纤链路层协同响应的生存性优化机制，造成了即便是大型的 IP 网络，也尽可能采用轻载的方式进行运维。此外，这几种承载方式的组网范围也根据传输距离的不同进行了较为严密的适用性界定。

由于 IP 网络的业务流量具有突发性，业务模式具有多变性，业务保障具有局限性，这对于大量业务的汇聚传导和疏通调度将会造成严重的制约。IP 网络中各级路由器拓扑数据的精确程度、网络规模的拓扑形式直接影响着被预置为备份切换的路径选择的可用度与响应度。对于 IP 网络的保护恢复主要涉及机制响应后在最迅速的切换时限转迁信息业务，同时有效保证切换后的信息业务质量，如能合理布局，其节点故障和链路故障是可以利用节点层面的可靠性技术和网络层面的保护恢复技术来提高 IP 的可靠性与安全性。

1.2 光网络的发展现状

传统光网络通常按照职能分工划分为管理界面和传送界面，各个界面均依托于数据通信网进行信息通信的“无缝连接”。管理界面定位于高端层次，负责对全网进行预设、监测、配置与优化等相关职能的统筹性工作；传送界面定位于低端层次，面向链路层依照交换方式的特性进行信息业务的传导，是一种多类传输样式的集合。

ASON（Automatically Switched Optical Network）的简称为自动交换光网络，是指在信令通信网控制下实现智能光网络连接自动交换功能，具有网络资源按需动态配置能力的光传送网络。这种新型光网络在网络结构上添加了一个控制界面，强化了光网络动态交换连接的自

适应效能，这也是区别于传统光网络的一个最重要的特征。ASON 的控制界面涵盖了一系列实时的信令与协议系统，承接了管理界面和传送界面的智能调度按需配置，依托于数据通信网中的信令通信网（SCN）对传送界面实施操作。事实上，ASON 的控制界面所承担的业务职能是传统光网络管理界面的操作职能。然而，ASON 的管理界面的职能没有因此而被弱化，而是管理职能变得细化，行为趋于简捷，权限更加高端。具体来说，控制界面表面上是从管理界面的职能剥离，而从实际层次分布的职能分析，管理界面承担了筹划的作用，控制界面则履行着执行的工作。甚者，在动态变化的资源配置过程中，控制界面亦可根据实体需求合理地进行智能优化和兼顾组织。

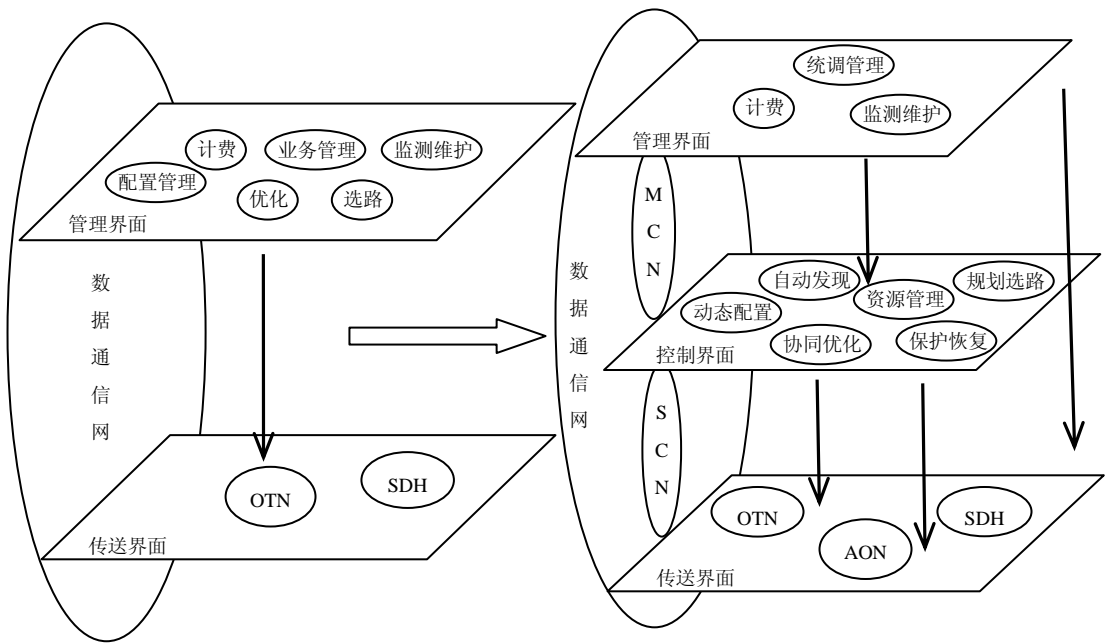


图 1 传统光网络向 ASON 的结构演进

在 ASON 的网络布局中，管理界面仍然是整个系统的中枢核心，对网络整体的架构负有决策性职能。它凭借着数据通信网这个信息平台，通过管理通信网（MCN）实施对全网的规划，既担负了其他界面间的协调与配合，又可直接干预对网络全局进行资源配置和连接管理，与控制界面互为补充，互为优化，在集中控制的点击式光通道配置中发挥着积极作用。

1.3 新型光网络承载IP的融合发展

随着全球 Internet 信息流量的与日俱增，亚洲地区信息业务的需求也呈现突飞猛进的状态。从 CNNIC 最新公布的中国互联网发展报告显示，至 2009 年上半年统计，中国的互联网用户已超过 3.39 亿，其中宽带网民的规模达到了总网民数的 94.3%。根据东西部区域之间，城乡之间对宽带的市场需求，智能升级原有带宽，提供融合式全业务，强调高性能、差异化宽带服务的“光进铜退”时代已成为网络发展一把重要的时间标尺。

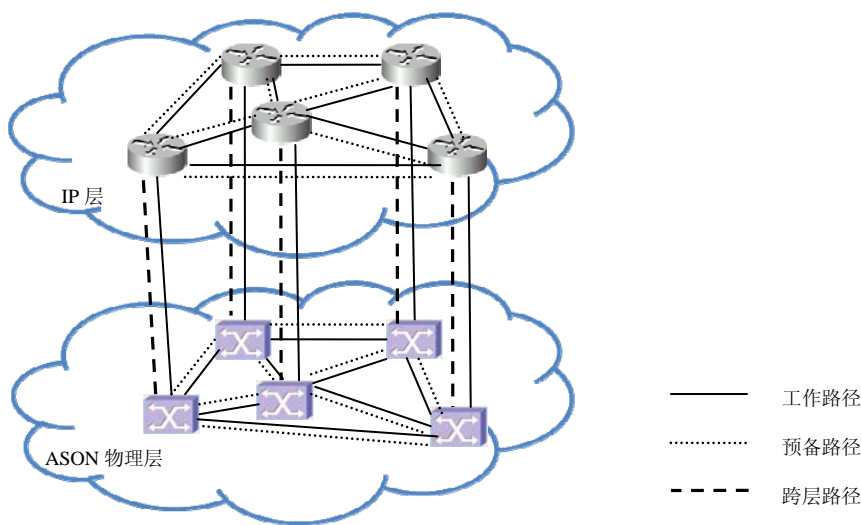


图2 IP Over ASON 网络架构

随着网络应用的推广，网络结构的复杂性给网络自身的可扩展性和生存性等一系列评估指标都带来了新的研究难题和发展空间。传统光网络依靠 SDH 环形链路链接简单实现了业务的保护恢复，然而，现今的光网络结构更为复杂严密，安全性动态配置显得极其重要，单一依靠高端网管实现全程监测与维护是很难达到实时调度和运通快捷的要求。在新型光网络承载 IP 融合建网中，最为突出的安全性特点就是光层恢复机制中光层故障对于 IP 层的透明性，这样使得即便是在光路阻断或是波分复用失效的状况下仍不影响 IP 业务的选路与链接。光传输系统以往主要集中对设备的管理，而对业务的及时管理甚少，在 IP Over ASON 的模式中，重点建设方向将放在实现业务的动态支持上。

IP Over ASON 的网络架构中主要还是通过主路径保障信息业务，当网络出现链路故障时，将会立即启用故障链路段的预备路径，把主路径的业务切换至接替路径上，或是按照优先级协议实现业务转移，对额外业务的链路进行按需调度配置，服务保障均衡兼顾，由保护和恢复的结合方式来保持信息传导的连续性和及时性。当网络出现节点故障时，IP 层通常只能通过自身进行排查，但在新的网络架构中，将会引入一种智能协同机制，对节点的连通分布域进行描述，把可用节点的可达链路进行判定和收敛，运用跨层路径的可控链路资源来弥补节点的局部失效，同时控制域也在实施对失效节点的修复，这样有利于提高信息业务的稳定性。

IP Over ASON 的管理是朝着综合化发展，力求将集成化统筹与分布式智能相结合，满足面向运营者的维护管理需求和面向用户的动态业务需求，实现多区域多样式多层次化的分布式协同管理。在新型光网络承载 IP 的革命性创新进程中，必将展现一种网络架构、业务支撑、网络管理和运营维护等基础体系的新模式。

2 应用可行性分析

2.1 社会可行性

从社会信息网络的主流发展方向而言，网络的经济型融合已是一种趋势。网络体系架构

的重组,“三网”全业务的融合,以及统一通信的问世,皆是社会信息市场的需求定位的客观反映。现代高新技术的不断更新也是推动网络延续改造升级的原始动力,ALL-IP 融合式网络已随着信息业务链的拓展愈发受到社会信息市场的追捧。

面向新时期军队信息化建设的长期布局,应充分引入成熟的科研技术对军用网络系统加以充实。根据军用网络建设应具有完整性、适应性和持久性的特点,亦可适时加强对带有军队特色核心技术的攻关,毕竟现有军用网络的安全性依然是目前世界军工行业研究的前沿。

2.2 经济可行性

通过引入ASON为IP网提供保护恢复,既可提高IP网的资源利用率,又能节省IP网的初始成本,但是ASON设备的部署从另一方面也增设了网络总体布局成本的新开支,节省开支与新增开支的经济逆差是否符合改造建设的投资规律,经济价值的评判恰恰取决于能否通过新技术支持下的新体系网络的运维效能来检验其高稳定性、强适应性的预期设想。

从经济开销的横向层次比较,首先需要考虑的就是设备的标准费用。由于现有光网络中均普遍使用环网实施保护恢复,不同厂家因为产品性能的迥异也存在着设备差价,但随着未来Mesh网络的建设与应用,成熟技术的研发共享,以及信息市场供求均衡的良性循环,各设备商的成品价格都将会趋向于平稳,这些因素的发展都是有利于保护建设方的长效投资。

军用网络建设关键在于通过经济支持从而引入成熟技术应用到全域网络的换代改造中,其最初的出发点和最终的落脚点都是要定位于提升信息化水平,具备适应在复杂环境中信息作战的能力。军用网络的拓展并不是一般运动的简单起始,而是一个依序渐进,长效规划的循环过程。通过经济的长期运筹,军用网络是可以支撑信息作战为军队赢得主动权。

2.3 技术可行性

各个层面的网络在应用范围上因为是相对独立组网,所以ASON的技术升级丝毫不影响与传统网络间的互联互通,还同以往传统网络的技术实现了兼容,从而开凿了一条自主可扩的并行发展路线。目前,IP Over ASON的承载方式主要是直接承载和部分承载,直接承载方式尽管对于技术的要求不是太严格,然而资源利用率却不是很高,经济投资的收益很难符合原始预期要求;部分承载方式对于互联接口和信令等相关技术有较为严格的框架限制,并且还要求路由器具备建立连接的响应机制,但是其经济性较好,可节省一定成品资源,倘若在网络中采用重载方式,资源利用率可达到80%,应用前景相对乐观。

ASON本身就是光网络和IP融合的产物,尤其是基于IP的协议(如OSPF-TE、RSVP-TE等)逐渐被智能光网络作为控制界面所采用。可以说,IP和ASON的融合承载本来就存在着一定意义上技术的相似与整合,这是一条网络优化必经的革新之路。

军用网络的生存性将直接决定军队信息化作战的总体效能,结合多种策略方案优化运用应是满足不同层次信息化作战需求的可行途径。在军用ASON融合建网的规划中,管理域,控制域,传送域亦可看作任务分工相对独立的自治域,但从安全性的监控角度来看,它们存在着一定联系的承接。通过各界面协同安全协议的修缮,逐步形成适用于军用网络的安全性机制,其间各域的协同策略是IP Over ASON模式一项实现资源利用率高效使用的有效措施。

3 应用生存性策略

3.1 分时策略管理

保护机制是根据各种业务类型的优先级而进行保护线路配置的网络拓扑设计，常用的保护方式有自动线路保护倒换、双归、自愈环的保护和网状网的保护等。这些保护方式均是在网络规划时所保护区域的链路进行预先铺设，在故障时实现链路对称切换的自动化响应，是一种预置的运维策略。恢复机制是根据各种业务类型的优先级而进行空闲容量征用的网络链路调度，预置重路由与实时重路由均属于恢复机制的范畴。预置是由网管系统根据信息业务线路的保障需求预先对其设置故障线路的计划转换，是一种静态的策略管理；而实时是基于计算机软件的算法功能对已出现的故障线路状态进行快速及时的链路统计分析，以更新现实可用的网络拓扑，对已损的信息业务实施倒换，是一种动态的策略管理。

一方面，在网络规划建设之初，就要根据信息业务的优先级特点，依照“预设为主，策略为辅”的原则，采取“动”与“静”相结合的策略管理，既能保障高等级用户的业务流畅，又可提高网络即时策略管理的成功率，发挥灵活度。另一方面，在一定的时域里，信息业务的传输时间段是可以有效进行针对性分级利用。在不同时差域内，可以有效利用时差间隙为不同需求的业务提供相应的动态带宽链路资源，实施业务的时差性分级保障。

基于军用通信保障优先级标准制定出对应的业务保障优先级，通过其间关系的映射，体现制定策略管理的安全系数。军用网络采用分时管理的优势在于充分利用各时差间信息业务的保障需求差异适时地提供服务，从而满足网络服务性能守衡的运维条件。在网络资源有余间隔，可进行局部硬件改造和软件升级等额外业务，若在网络资源稀缺时，就必须聚合可用链路资源按照优先级规定逐次予以保障，随着各优先级用户保障需求的变化，分时策略管理也应灵活调整，加以实时的动态配置，保持军用网络各用户安全系数的均衡水平。

3.2 分域策略管理

在全域的网络布局中，物理链路相对于逻辑链路而言，实际的链路部署要更多，常留有备份用以保护和恢复，为了预防及保障应急情况中信息业务的即时通信。安全性策略也大多采用资源备份的方式，这样有助于故障出现时的快速转换，保证信息流动交互的持续性。

及时的链路监控是实施策略管理的前提，只有通过正确的检测，才能发现受损链路或节点的故障，即可发挥分域策略管理的作用。区域性的路由策略选取需要在两种情况之下做出决策，其中有信息业务在地理区域的大量汇聚，造成信息传导阻塞所引发的暂时性逻辑断路，这需要进行业务分流，使域内信息业务负载均衡后，逻辑断路问题便可自然迎刃而解。另一种是传输链路因自然性灾害或人为活动而造成的持久性物理断路，这需要对信息业务进行成批转移，首先对域内进行可达路由收敛和链路更新统计，然后对受损链路的业务进行分域切换，逐次逐批地修复信息业务。在网络拓扑重构过程中要尽量避免域内的多个路由器重复交换相同信息，或各个自治系统执行相冲突的路由策略造成协议的统一实效，如果存在这些不确定因素而引发路由振荡，对域内的保护恢复协同策略将会产生影响。

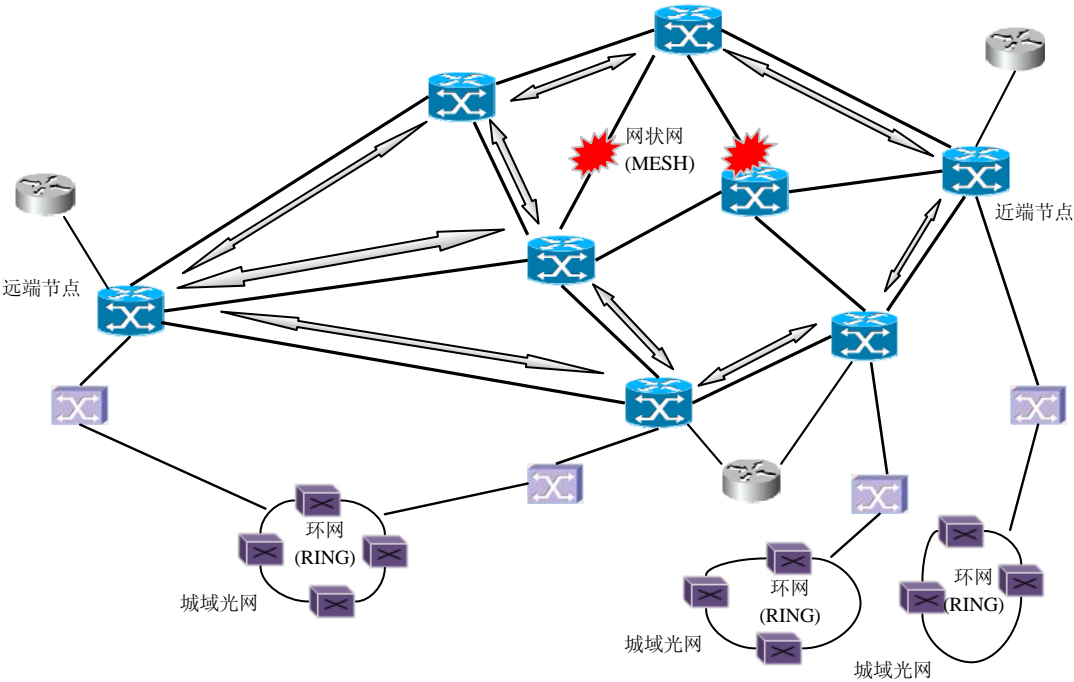


图 3 信息业务在网络拓扑中的策略管理

在军用网络的应用体系中，全方位容灾备份方案对于系统整体而言显得更加重要。由于系统的扩容，信息交换量也将增多，全 IP 通信设备对系统稳定性和可靠性要求则更高，这就需要在各个层次做好全方位的资源备份。尤其在复杂的战场环境中，一些不可预测的因素和战争进程的时变所带动保障资源的调整都将会引发网络的局部失控，在对待关键性信息业务的支持上要本着“就近原则”施以恰当的分域策略管理，最大限度激活网络的整体效能。

结束语

如何在有多条件约束的复杂环境下克服非连续工作带宽级联限制，合理提升资源利用率，以及 ASON 组播智能可控技术的应用对 IP Over ASON 的容灾抗毁机制的促发策略等均是今后信息通信领域仍需深入探讨的话题。经历了全 IP 网络发展，特别是以 IP 协议为基础的业务在不同网络的互联互通，将有效改善电信级的网络能力保障、网络可靠性、组网灵活性以及网络服务质量等系列的硬性指标，为 ALL-IP 通用电信级平台的融合建网打下了良好的基础。因此，下一代网络承载和控制分离、控制和业务分离的思想与以软交换为基础的 IMS 移动网络的演进思路相得益彰。

参 考 文 献

[1] 徐云斌, 张海懿. ASON 网络运营维护技术研究[J]. 光通信技术, 2009 (1), 19-21

[2] 李 健, 邓 宇, 刘海玉等编著. ASON 网络互联[M]. 北 京: 人民邮电出版社, 2008.07

[3] 周荣生, 杜春生. 对 ASON 承载 IP 网的研究[J]. 电信科学, 2007 (3), 21-26

[4] David Benjamin, Richard Trudel, Stephen Shew. Optical Services Over The Intelligent Optical Network[J].

IEEE Communication Magazine, 2001.09, 73-78

- [5] Crochat O et al. Protection Interoperability for WDM Optical Networks[J]. IEEE/ACM Transactions on Networking, 2000.08, 384-395
- [6] PERELLO J. Transport Plane Resource Discovery Mechanisms For ASON/GMPLS Meshed Transport Networks[R]. Lecture Notes in Computer Science, 2007.06, 221-228
- [7] P.H.Ho et al. A Framework for Service-Guaranteed Shared Protection in WDM Mesh Networks[J]. IEEE Communication Magazine, 2002.08, 97-103
- [8] Ling Wang, Peida Ye. The Optimal Design of Logical Topology with oS Constraints in IP over WDM Network[C]. Proceedings of International Conference on Communication Technology, Beijing, 2003, 126-129
- [9] Chao Wang, Yanhe Li, Xiaoping Zheng, et al. Study on a novel traffic engineering model for IP over ASON network[C]. Proceedings of Asia-pacific Optical Communications, Beijing, 2004, 270-279

作者简介

任志宏（1984 - ），男，贵州省贵阳市人，重庆通信学院在读硕士研究生，主要研究方向通信对抗和网络安全。

谢永强（1972 - ），男，安徽省黄山市人，高级工程师，中国电子设备系统工程公司网络中心主任，研究领域网络信息安全与安全操作系统

赵广松（1984 - ），男，江苏省盐城市人，解放军理工大学在读研究生，主要研究方向网络信息安全

IPTV业务在现有宽带网络中的实现

赵 怡

(重庆电子工程职业学院 重庆市 401147)

摘 要: IPTV (Internet Protocol Television) 业务不同于传统的视频通信业务和广播电视业务, 它代表了基于 IP 网络的新型传播视频业务的发展方向。由于 IPTV 本身是一种全新的业务形式, 其业务特性和发展方向还有待进一步明确。所以, 目前国内还未全方位大面积地开展此项业务。要最终构建全国性的能满足商用运营需求的 IPTV 业务网络就必须对 IPTV 技术开展循序渐进的研究。本文就逐一对 IPTV 的技术和业务现状、主流营运商宽带网络现状以及在现有宽带网络中承载 IPTV 需要解决好的几个关键技术进行了阐述和探讨。

关键词: IPTV; 组播; QoS; 宽带接入网络

1 IPTV概述

从上世纪 90 年代开始, IPTV 就受到了越来越多人的关注。IPTV 业务是指以 IP 为传送技术, 以 TV 作为媒体终端, 以交互式音视频服务为主体的崭新业务集合体。它利用宽带网络, 集视频编解码、流媒体、宽带通信、数字版权等多种技术于一身, 向用户提供交互式音视频服务。目前, 随着宽带网络技术、视频编码技术迅速发展及业务运营经验的积累, IPTV 业务正逐渐走向实际运营阶段, 成为电信运营商的主流宽带业务。

IPTV 的业务形式多样, 但主要是直播、点播及两者的变体。

1) 直播

直播是 IPTV 业务的基本业务形式之一, 对用户而言此种业务如同传统频道电视, 频道切换和频道选择通过屏幕菜单形式实现, 丰富了用户的收视频道; 对运营商而言, 直播业务是吸引传统电视用户的有效手段, 其运营关键是频道特色; 从技术实现角度看, 此种业务一般采用 IP 组播技术在 IP 网络上传送 TV 节目信息。直播节目内容首先推送到 IPTV 传送网内, 由传送网内组播源通过 IPTV 传送网组播发送到汇聚层边缘业务接入控制点 (BRAS/AR), 再由业务接入控制点 (POP 点) 通过接入层提供给用户。因此要求承载网络支持 IP 组播和 IP QoS。

2) 点播

点播是 IPTV 业务的另一种基本业务形式, 用户通过屏幕菜单选择播放内容, 且按内容支付费用。它彻底颠覆了用户的收视习惯, 变被动收视到主动点播, 把播放内容的选择权交给了用户, 极大地满足了用户的个性化需求, 是完全有别于传统电视的崭新业务形式。

根据收费模式、目标客户群体、内容种类的不同, 这种基本业务形式又可以演化出多种变异形式。如: PVR (Private Video Recorder) --个人电视录播; PPV (Pay Per View) --按次付费点播; TSTV (Time-shifted Television) --时移电视; NVoD (Near Video on Demand) --准

VoD 或轮播。还有电视购物，互动游戏等等。
从技术实现角度看，此种业务一般采用 IP 单播技术在 IP 网络上传送 TV 节目信息。

2 IPTV业务模型及总体结构

整个 IPTV 系统是一个庞大复杂的系统，汇聚了视频处理、数据通信、数据传输、业务运营管理等多个领域的技术，贯穿了视频源网络、视频服务网络、宽带接入网络、机顶盒设备等多个网络和设备。IPTV 数据承载网主要包括两部分，一是 IPTV 视频服务网络的承载网——IPTV 传送网络；二是 IPTV 宽带接入网的数据承载部分。配套数据设备主要用在 IPTV 传送网络，宽带接入网部分一般是利用运营商原有网络。

一个典型的 IPTV 系统网络模型如图 1 所示。

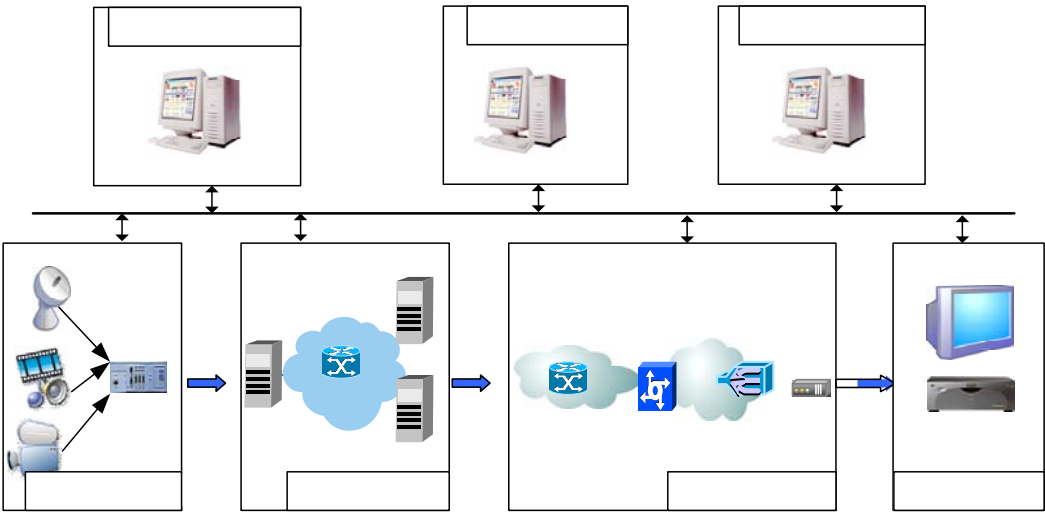


图 1 IPTV 系统网络模型

其中：

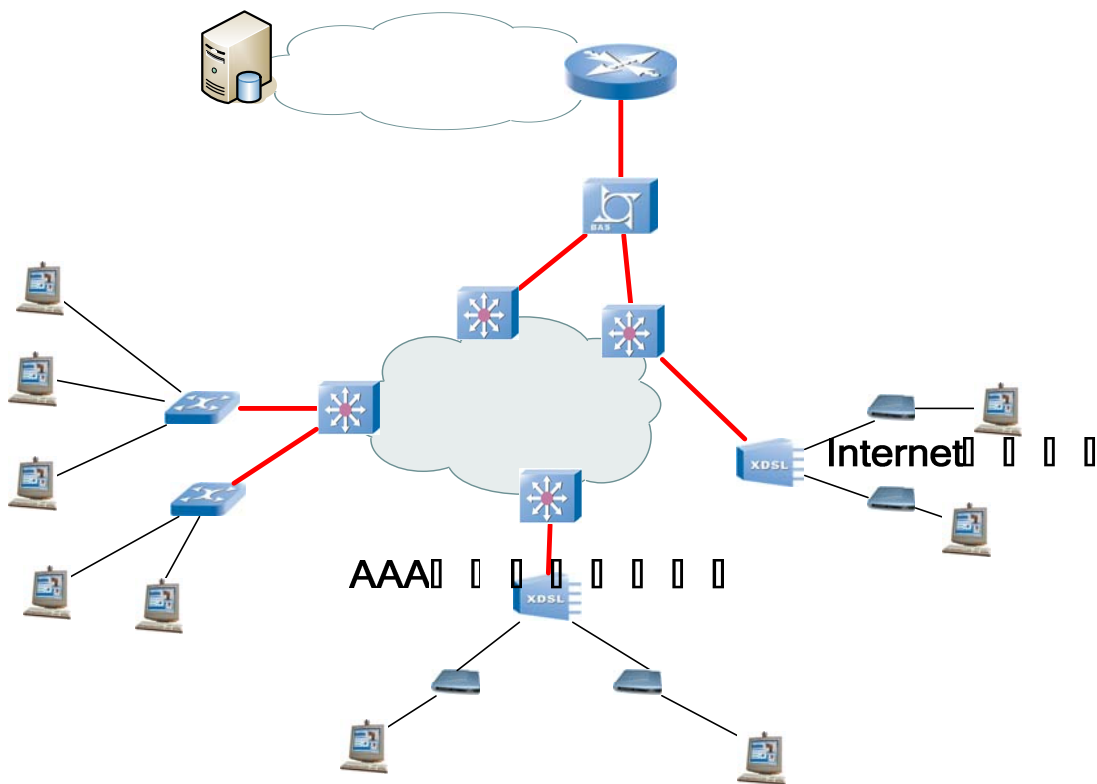
IPTV 视频服务网络是位于视频源系统和宽带接入网之间一段网络，它完成视频数据的导入、存储、分发和服务等功能。视频服务网络的基本原理是把视频内容推送到网络边缘，为用户就近提供服务，从而有效提高了服务质量，降低了骨干网络的传输压力，为 IPTV 业务规模应用提供了基础。**IPTV 传送网络**主要承载单播类 IPTV 业务以及组播类 IPTV 业务。

IPTV 宽带接入网从视频服务网络到用户终端的一段网络，配合运营管理网实现用户宽带上网接入认证管理、视频组播组加入、离开控制功能，并将用户需要的视频流发给用户，为用户接入 IPTV 业务提供有 QoS 保证和传输通道。

IPTV 视频服务网络不直接面对大量接入用户群，营运商可就用户的分布情况新建一套数据网络将各种视频流推送到各地的汇聚层。宽带接入网负责数万甚至数百万分布于城市、农村各地的最终用户接入，如果为此新建一套专用于 IPTV 业务的宽带接入网络，将会进行重复的巨额投资建设。利用营运商现有的宽带上网接入网络平台一并解决用户上网和 IPTV 业务接入成为必须解决的问题。

3 目前主流营运商宽带网络现状

目前各运营商宽带数据网络总体结构如图 2 所示。



密度远小于 DSLAM 设备，因此 LAN 接入方式一般适用于新建小区。

两种接入方式各有优缺点，目前主流运营商两种接入方式并存，相互补充（近年来各大运营商也开始新建 PON（EPON、GPON 等）接入网络，发展十分迅速）。普通用户上网一般采用 PPPOE 拨号方式上网，两种接入方式都是一样。用户端发起 PPPOE 呼叫，BRAS 提供 PPPOE 服务，用户端拨号软件与 BRAS 在反复交互协商后用户端发送合法的用户名和密码，BRAS 上送 AAA 服务器对用户进行账号/密码的认证，认证通过后进行授权，给用户分配相应的网络资源（上网的 IP、DNS、路由等），用户上网后，AAA 服务器对用户进行计费管理。用户下线后释放相应的网络资源。

4 在主流营运商宽带接入网络中实现IPTV业务的承载

前文已叙述 IPTV 业务流从节目源传送到用户终端时需要经过 IPTV 传送网络（视频服务网络）和用户宽带接入网络。传送网络可以新建一套独立于互联网系统的网络，传送网络将 IPTV 的视频流传送到各传送网络的各边缘节点（POP 点），边缘节点将视频流注入传统的宽带接入网络中，利用传统的宽带接入网络将视频流传送到最终用户。这样既保护了原有宽带汇聚和接入网络的投资，无需重复建设，又能够快速部署 IPTV 全网业务。目前主流营运商均采用此种方案来解决 IPTV 视频流传送到最终用户的问题。典型的业务网络如图 3 示。

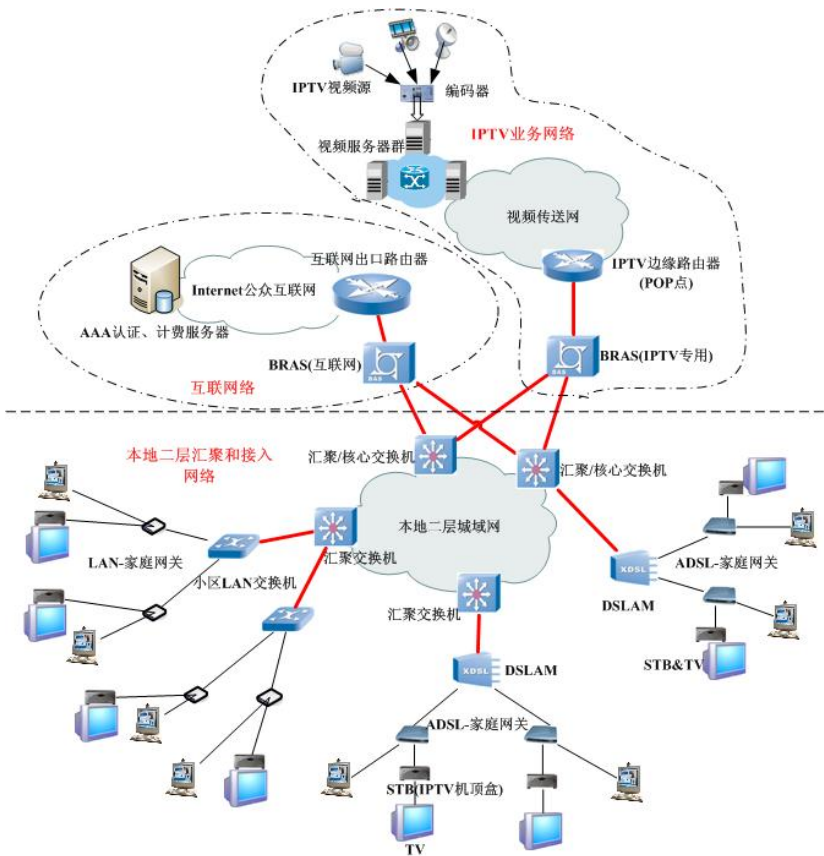


图 3 典型 IPTV 业务/互联网业务网络图

由图可知，上层互联网络与 IPTV 业务网络相互独立，也就保证了上层互联网与 IPTV 业务网络的业务在骨干层面相互独立，互不干扰。对于 IPTV 这种对带宽和时延要求极为苛刻的业务来说是最为理想的方式。（注：在业务量不太大的情况下，为节省投资，IPTV 用 BRAS 有时也和互联网 BRAS 共用，采用不同域来区分，BRAS 根据不同域上联到各自骨干网络）。

本地二层汇聚和接入网络则采用同一张网络，在核心交换机上分别与互联网骨干层和 IPTV 业务网络对接（通过不同的 VLAN 进行业务的区分，如一段 VLAN 用于互联网、一段 VLAN 用于 IPTV 等），那么互联网业务和 IPTV 业务在本地汇聚接入网络上混合传输的。那么，这两种业务在本地汇聚/接入网络上如何区分、业务质量如何进行保障、如何进行用户的接入是本地网络需要解决的问题。

5 现有宽带网络中承载IPTV涉及的几个关键技术

本地二层汇聚接入网络在承载传统的互联网业务基础上，还需要承载新的 IPTV 业务，因 IPTV 中直播业务采用组播技术，在整个传送网络中组播复制可在不同控制点实现，因此，整个网络就需要解决用户接入、业务区分、业务质量保证、组播复制等多个问题。

1) IPTV 接入网组播技术

直播业务流主要通过 IPTV 传送网传送至接入网，最终由接入网送抵用户终端，完成组播业务数据平面传送工作。

根据组播复制/控制点的不同，接入网的组播大致可以分为以下三种形式：基于 BRAS 的组播复制方式、基于汇聚交换机（组播交换机）的组播复制方式、基于 DSLAM/二层交换机的组播复制方式。

① 基于 BRAS 的组播复制方式

本实现方式用户 STB（IPTV 机顶盒）通过使用 PPPOE 或者 IPOE 方式接入，与 BRAS 之间建立 PPPOE 或者 IPOE 通道，BRAS 终结 STB 的 IGMP 报文，由 BRAS 负责实现用户 STB 的组播复制，将组播报文复制在 STB 相应的 PPPOE 或者 IPOE 通道内，具体实现方式如图 4 所示。

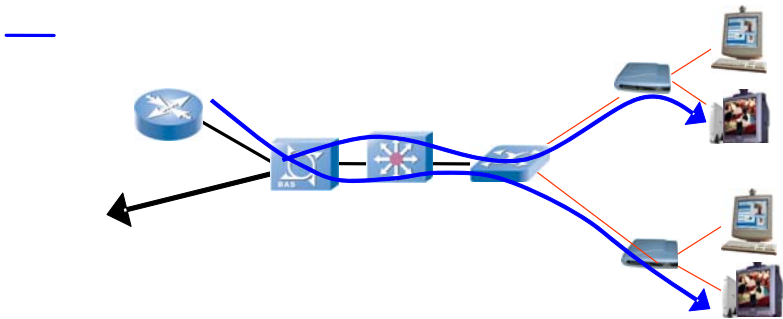


图 4 基于 BRAS 的组播复制方式

该实现方式适合单边缘业务接入+用户单通道接入、单边缘业务接入+用户多通道接入的接入方式，PC 采用 PPPOE 方式接入，STB 采用 PPPOE 或者 IPOE 方式接入，PC 和 STB 即可以共用一条 PVC/VLAN，也可以分别做单独配置。本实现方式不需对现网做太大改造，

适合采用“集成模式”组网情况，只要求对宽带计费后台进行少量的改动即可。但该实现方式中，BRAS 面向用户 STB 复制 IPTV 组播业务，面向用户的组播复制点是 BRAS，对 BRAS 的下连带宽要求很高，不适合大规模 IPTV 的组网。建议在 IPTV 业务开展初期使用。

② 基于汇聚交换机（组播交换机）的组播复制方式

本实现方式用户 STB 可以采用多种接入方式，PPPOE 或者 IPOE，但采用 PPPOE 接入时，STB 必须支持双栈，能够发送基于 IPOE 封装的 IGMP 报文，汇聚交换机终结 STB 的 IGMP 报文，负责将组播 M-VLAN 的 IPTV 直播业务跨 VLAN 复制给用户，图 5 给出了在单边缘业务接入情况下具体的实现方式。

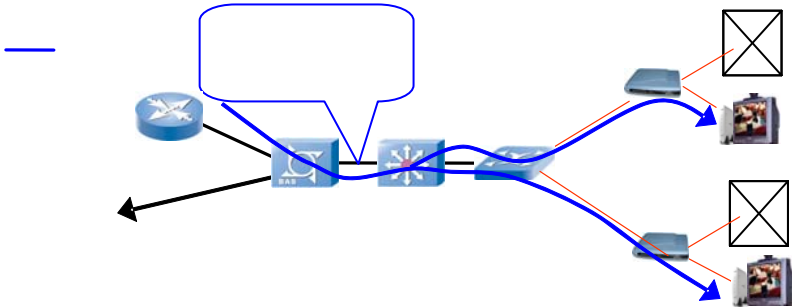


图 5 基于汇聚交换机的复制方式

该实现方式适合所有接入方式，可以是单/多边缘业务接入、用户单/多通道接入的组合，但当 STB 使用采用 PPPOE 接入时，STB 必须支持双栈，能够发送基于 IPOE 封装的 IGMP 报文；汇聚交换机至 IPTV 业务控制点之间的直播业务采用组播 M-VLAN 承载，汇聚交换机具备 IGMP Proxy 功能，可以采用主动静态下拉或者动态下拉的方式将 IPTV 组播业务通过组播 M-VLAN 送抵至汇聚交换机，然后按需跨 VLAN 复制给用户；所谓主动静态下拉就是指不管有没有用户需要组播流，汇聚交换机均主动向上行发送组播加入报文进行引流；所谓动态下拉是指只有当有第一个用户组播加入时，才进行引流，后续用户不再进行引流，当所有用户组播均离开时，汇聚交换机发送组播离开消息切断组播流，从而实现“按需引流，一次引流，多用户应用”的目的。该实现方式对现网改造不大，而且支持所有接入方式，缓解了 BRAS 的接入压力，但相应的汇聚交换机与 DSLAM/二层交换机之间，原有基于 BRAS 组播复制的带宽压力没有改善，而且还要考虑控制用户组播接入的问题，适合 IPTV 业务开展的过渡阶段。

③ 基于 DSLAM/二层（接入）交换机的组播复制方式

本实现方式用户 STB 可以采用多种接入方式，PPPOE 或者 IPOE，但采用 PPPOE 接入时，STB 必须支持双栈，能够发送基于 IPOE 封装的 IGMP 报文，DSLAM（二层交换机，以下出现 DSLAM 的情况同样适用于交换机）终结 STB 的 IGMP 报文，负责将组播 M-VLAN 的 IPTV 直播业务按接入 PVC/端口复制给用户，图 6 给出了在单边缘业务接入情况下具体的实现方式。

该实现方式适合所有接入方式，可以是单/多边缘业务接入、用户单/多通道接入的组合，但当 STB 使用采用 PPPOE 接入时，STB 必须支持双栈，能够发送基于 IPOE 封装的 IGMP 报文；DSLAM 至 IPTV 业务控制点之间的直播业务采用组播 M-VLAN 承载，DSLAM 具备 IGMP Proxy 功能，可以采用主动静态下拉或者动态下拉的方式将 IPTV 组播业务通过组播 M-VLAN 送抵至 DSLAM，然后按需复制给用户。该实现方式现网做改造较大，对原有的不支持组播复制的 DSLAM 均要更换或升级，支持所有接入方式，缓解了 DSLAM 以上的组播复制压力，

但还要考虑控制用户组播接入的问题，在大规模 IPTV 放号阶段，DSLAM 作为组播复制/控制点，是较好的选择。

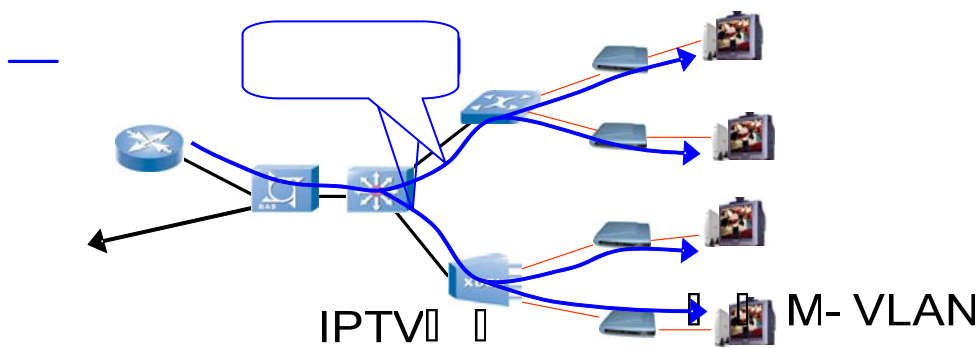


图 6 基于 DSLAM/二层交换机的复制方式

2) 现有宽带汇聚接入网络需解决的几个技术问题

根据前文所述，IPTV 在本地汇聚接入网络中有几种不同的组播复制方式，每种不同的方式对现有宽带接入网络的要求会有所不一致。就几种情况综合而言，需要解决好以下几个主要技术问题。

① 用户端口定位

随着 TCP/IP 协议的盛行并在网络中占据主导地位，宽带接入网经过一段时间的高速发展，宽带用户数量已经增加到一定的规模，一个本地网络接入用户可达几万到数十万，在 IP 环境下，使用 VLAN 来标示用户的物理位置信息。但是，由于 802.1q 协议本身的限制，VLAN 的数量被限制在 4094 个，这就使得传统 VLAN 区分的方法成了很大的限制。同时网络新业务也在不断发展，对于新业务，在接入层对于其接入的合法性的鉴别有了新的要求。在这种情况下，如何精确的鉴别一个用户及业务接入的合法性，实现用户的可管理性，对运营商有着相当重要意义。因此，需要新的技术来解决用户物理位置信息，从而完成用户物理信息和端口的绑定，保护用户和运营商的合法权益。

因此 VBAS、PPPOE+、DHCP option82 以及 SVLAN (stack VLAN) /QINQ 等用户端口定位技术应运而生。目前这些技术在主流运营商中被使用（特别是大规模本地网络）。

SVLAN/QINQ 技术（也被称为双层 vlan）是一种实现较简单可靠，设备间无需协议对接的端口定位技术，它在传统 802.1q 协议报文基础上，在其外层再增加一层 802.1q VLAN。这样，双层 VLAN 的范围被扩充到了 4094*4094 约 1600 万个，完全可满足本地网络数万到数十万用户的端口定位。目前大多运营商首选此定位技术。在布置 SVLAN 的接入网络中，根据双层 VLAN 实现位置，需要对应的设备支持 SVLAN/QINQ，如 DSLAM 设备、交换机设备、BRAS 设备（BRAS 作为双层 VLAN 的终结设备，必须支持）。

② 组播支持

根据 IPTV 组播复制点的不同，相应的组播复制设备必须支持组播功能和性能支持。如 DSLAM 设备、交换机设备、BRAS 设备。

③ QoS 保证

宽带接入网络中各个设备上均为用户传送互联网业务及 IPTV 业务流，通过不同 VLAN 或 PVC（DSLAM 设备接入用 PVC）承载。两种业务中间走的物理通道相同，但是两种业务

对服务质量要求不一样,传统互联网业务要求较低,IPTV 业务传输的视频数据,实时性很强,对 QoS 要求很高。

IPTV 业务对 QoS 的要求反映到承载网络层面,即为对带宽、网络时延、时延抖动、丢包和可靠性的要求。足够的带宽是提供良好业务质量的基本条件,时延、抖动、丢包三个网络参数是最基本的网络层 QoS 指标;另外,网络还要提供有效的可靠性,在出现链路或节点故障的情况下,要及时检测并采取相应的措施,来保证业务的不间断或迅速重启动。

在 IPTV 体系框架下,QoS 的保障不仅仅是对网络设备提出要求,它需要的是提供一个端到端的完整解决方案,即从一个终端的应用层到另一个终端的应用层的整个流程中,各个环节均应具备 QoS 保障。

因此本地汇聚接入网络要有针对不同业务提供不同 QoS 的能力。这些设备包括 BRAS、交换机、DSLAM。DSLAM 设备用户端口要支持多 PVC 方式,不同的 PVC 走不同的业务。

④ 用户终端接入

与传统的宽带上网接入比较,最终用户接入网络时需采用 LAN 或 DSL 的家庭网关,这种家庭网关上行支持多 VLAN 或多 PVC 方式与局端设备对接走不同的业务。下联多接入口分别接上网用 PC 和 IPTV 的 STB 设备。

⑤ 接入网络带宽和要求

IPTV 一个普通视频流约占 2M 的带宽,高清视频流约占 10M 左右。因此,需给最终用户提供不低于 2M (或 10M,高清视频)的下行接入带宽,考虑到用户还有到互联网的流量,一般给用户的接入带宽不低于 3-4M (如提供高清则需 12M 以上带宽)。本地二层网络根据最大同时在线观看 IPTV 用户*2M (如 500 个在线点播用户需 1G 的带宽)可大致估算宽带接入网络主干电路上需预留或规划给 IPTV 业务的带宽。IPTV 业务直播节目采用 UDP 传送,无重传机制。因此对丢包率、时延及抖动要求也十分苛刻。

小结

利用现有的宽带接入网络解决 IPTV 业务的接入是各营运商需解决的问题。因 IPTV 业务的特点,对接入网络也有较为苛刻的要求,可能需要对原网络进行必要的升级和改造,以满足承载 IPTV 业务的要求。综合起来,本地汇聚接入网络需解决好组播能力、QoS 保障、网络带宽、接入线路质量、端口定位能力等方面的内容。这样,利用传统宽带汇聚接入网络同时实现互联网业务和 IPTV 业务的承载将成为必然。

参 考 文 献

- [1] 《中国 IPTV 产业动态》第 16 期 (2009 年 8 月号)
- [2] 《中国 IPTV 产业动态》第 21 期 (2010 年 01 月号)
- [3] 《中国 IPTV 产业发展研究报告》流媒体网 2009-02-17

作者简介

赵怡 女,1973 年 3 月,讲师 主要的专业研究领域为计算机网络、计算机信息管理和计算机信息安全

第 5 部分

系统集成技术

软件项目计划与跟踪

刘卫宏¹ 焦彦平²

(1 装备指挥技术学院 信息装备系 北京 101416;

2 装备指挥技术学院 科研部 北京 101416)

摘 要: 项目计划和项目跟踪与控制是 CMMI2 级的两个项目管理过程域。项目计划的主要目的是建立和维护那些定义项目活动的计划; 项目跟踪与控制的主要目的是做到对软件开发的实际过程有适当的可视性, 使管理者能在项目的软件过程实施明显偏离软件计划时采取有效的纠正措施。本文主要介绍了如何将这两个过程结合起来进行项目管理。

关键词: 项目计划; 项目跟踪与控制

The Software Project Planning and Monitoring

LIU Wei-hong¹ JIAO Yan-ping²

Abstract: The purpose of Project Planning (PP) is to establish and maintain plans that define project activities. The purpose of Project Monitoring and Control (PMC) is to provide an understanding of the project's progress so that appropriate corrective actions can be taken when the project's performance deviates significantly from the plan. This paper focuses on introducing how join these two processes to manage the project.

Keywords: Project Planning , Project Monitoring and Control

1 前言

项目计划的目的是为项目的开发制定合理的行动计划, 使项目的所有人员能够按计划完成工作。项目跟踪与控制的目的是通过定期的检查项目计划的各种指标, 了解项目的进展情况, 并在项目的进展情况与计划有较大偏差时, 及时地做出调整, 使项目回到正常的轨道。主要是对进度、费用、工作产品和工作量等的跟踪控制。

项目计划和跟踪主要由项目负责人管理。项目的计划和跟踪过程对项目的成败具有非常重要的意义。项目计划与项目跟踪是两个相辅相成的过程, 如果没有计划, 则谈不上项目的跟踪; 如果没有跟踪, 则项目的计划便得不到落实, 起不到应有的作用。所以若想对项目进行有效的管理, 必须将这两个过程结合起来。

2 项目计划

软件项目的开发计划是跟踪软件活动、传递软件状态和修订软件计划的基础。管理者监

控软件活动，主要是通过软件产品完成时或在所选择的里程碑处，把软件开发过程中实际完成的规模、工作量、成本和进度表等与计划值相比较，确定项目的实际进展，并根据偏差，识别项目中存在的问题。当发现与软件项目开发计划偏离时，根据实际情况适当地修订软件开发计划，重新策划后续工作。

因此，可以看出项目的开发计划直接关系到项目的好坏、成败。项目计划制定得越详细、越准确，项目成功的几率就越大。然而，在项目的实际进展中，实际值与计划值总是存在着或多或少的偏差。于是就需要收集项目数据，并对偏差值进行详细的分析，找出导致偏差产生的原因。有了这些分析结果之后，就可以通知项目经理、阶段负责人、任务负责人以及质量保证人员等，采取适当的措施，例如修订项目计划等，保证项目在可控的范围内正常运行。

2.1 项目计划的主要活动

项目计划开始于对产品和项目需求的定义。

项目计划通常包括以下活动：

- 开发项目计划；
- 与相关人员进行适当地交流；
- 获取对计划的承诺；
- 维护计划。

项目计划的内容包括：

- 估计工作产品和任务的属性；
- 决定所需的资源；
- 商谈承诺；
- 产生一个进度表；
- 标识并分析项目风险等。

建立项目计划时，上面这些活动可能需要不断地反复。项目计划为执行和控制项目活动提供了基础。

为保证所有的技术和支持活动都在项目计划中有充分的体现，所有的相关人员都应该参与到生命周期各个阶段的计划过程中。

因为需求和承诺的变更、不准确的估计、纠正措施以及过程变更等原因，项目计划经常要随之修订。

在整个项目计划过程中，主要要完成的目标有三个：建立估计值、完成项目计划和获得高层领导对计划的承诺。

表 1 列出了与每个目标对应的实践活动。

表 1 项目计划的目标和活动

目标一	建立估计值
活动	1 估计项目范围
	2 建立工作产品和任务属性的估计值
	3 定义项目生命周期
	4 估计工作量和成本

目标二	开发项目计划
活动	1 建立预算和进度表
	2 标识项目风险
	3 计划数据管理
	4 计划项目资源
	5 计划需要的知识和技能
	6 计划涉及的相关人员
	7 建立项目计划
目标三	获得对计划的承诺
活动	1 评审影响项目的各种计划（例如：风险减轻计划、质量保证计划、配置管理计划等）
	2 调整工作和资源等级
	3 获取计划承诺

2.2 项目计划的设计

项目计划之所以称为“计划”，就因为它不是十全十美的，即使经验丰富的专业人员，制定的计划也总会存在一些偏差。而且，计划在开始的时候往往比较粗糙，对工作产品估计的粒度比较大；但是随着任务的细分，角色的分配，员工职责的分明，估计的会粒度越来越小，准确性也逐渐提高。

所以为了保证项目跟踪与控制的顺利实施，可以将项目计划设计为三个层次，分别为：项目总体计划，阶段计划和任务计划，其关系如图 1 所示。计划更改的时候，一定要保持各层次计划的一致性。总体计划会因阶段计划的更改而调整，任务计划受阶段计划的影响。

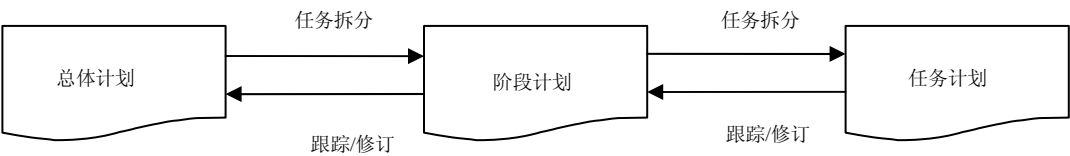


图 1 三个计划的关系

总体计划：项目经理根据用户需求制定总体计划，给出项目进行的主要阶段和各种需求。此计划需要经过评审通过后方可执行。总体计划要得到客户和高层领导的认可，并且必须通知到相关的人员，比如质量保证人员和配置管理人员等。

总体计划的主要内容包括：

- 项目生命周期阶段的划分；
- 项目规模；
- 进度表；
- 每个阶段主要活动的资源需求；
- 设定的里程碑；
- 可度量的质量数据；
- 总工作量、成本和人员。

阶段计划：项目经理、阶段负责人以及所有的参与人员共同制定阶段计划。阶段计划是

总体计划的任务分解。阶段计划要与总体计划保持一致，每项任务的估算得到完成人的认可，相关人员要清楚任务之间的依赖关系。

阶段计划的主要内容包括：

- 进度表；
- 任务、人员安排；
- 各个任务之间的依赖关系；
- 任务规模；
- 风险；
- 资源需求；
- 工作量、成本；
- 阶段结束的标准。

任务计划：根据阶段计划中的任务安排，每个人制定自己的任务计划。最小任务尽量只由一个人来完成，避免互相推诿或依赖。

任务计划的主要内容包括：

- 每天的任务；
- 提交的代码；
- 提交的文档；
- 任务完成率。

3 项目跟踪与控制

项目跟踪与控制是为了了解项目的实际进展情况而采取的一系列活动。例如了解员工任务的完成情况；了解项目在重要里程碑处的完成情况；了解整个项目计划的完成情况；根据跟踪结果进而了解人员士气、思想变化、员工能力等。跟踪主要是为了及时了解项目在实际进展中存在的问题，根据问题的严重程度，及时采取有效的解决措施，并使相关人员能实时了解项目的当前状态。

3.1 项目跟踪与控制的主要活动

已经归档的项目计划是监督活动、沟通状态和采取纠正措施的基础。

项目进度主要是在规定的里程碑、项目进度、工作结构图 WBS 中的控制等级处确定进度，把工作产品和任务的属性、工作量、成本、进度等的实际值与计划值进行比较。当性能明显偏离计划时，对进度适当的可视性可以促使管理者及时地采取纠正措施。如果重大偏移没有解决，那么项目将不能满足其规定的目标。

当实际状态明显偏离预期值时，适当地采取纠正措施。这些措施可能需要重新制定项目计划，包括修订原计划、建立新的协议，或者在当前计划中增加额外的缓解活动。

项目跟踪与控制的主要目标有两个：按照计划监控项目和管理纠正措施直到结束。

表 2 列出了与每个目标对应的实践活动。

表 2 项目跟踪与控制的目标与活动

目标一	按照计划监控项目
活动	1 监控项目计划的各项参数
	2 监控承诺的事项
	3 监控项目风险
	4 监控数据的管理
	5 监控相关人员的参与
	6 进行进度的评审
	7 进行里程碑的评审
目标二	管理纠正措施直到结束
活动	1 分析问题
	2 采取纠正措施
	3 管理纠正措施

3.2 项目跟踪与控制的设计

项目跟踪与控制可以验证项目开发计划是否正常实施，同时也可以证明该计划是否可以按时完成。因为跟踪可以检验计划，所以根据跟踪过程中发现的不当之处，及时对计划进行适当的改进。所以从这一点上，可以把计划和跟踪作为一个工作循环过程，通过这个循环的不断深入，计划将不断地得到改进，从而使得开发过程的可视性越来越强，项目管理越来越容易。

对应于项目计划的三层设计，项目跟踪与控制也设计为三个层次。如图 2 所示，三个层次的跟踪分别为：总体跟踪、阶段跟踪和任务跟踪。它们之间的关系如下所述。

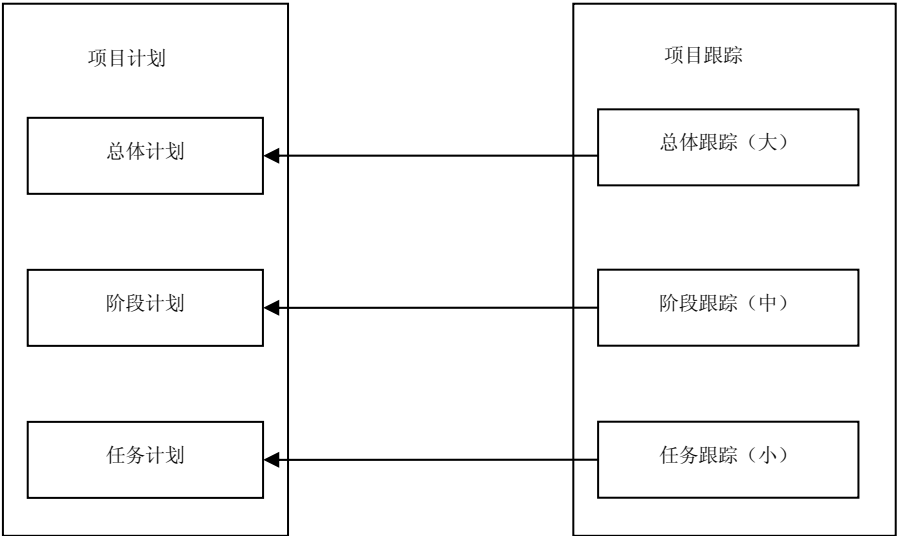


图 2 项目跟踪与计划的关系

● 总体跟踪

总体跟踪主要从项目的全局进行把握，对应到项目的总体计划。项目在开始时，一般是很难做出一个具体的计划来估计项目的细节工作，所以这时项目经理往往制定一个总体计划。该计划中一般不包含细节内容，只是大概估计项目的起始时间、总工作量、成本、规模等，并且包括相对简单的阶段计划。所以在这个层次上跟踪项目的实际起始时间、总工作量、总成本、规模、实际完成率等等。并将实际值与项目计划中的估计值进行比较，如果该偏差值超出定义的范围，则给项目经理、质量保证人员等发出警告，提醒其采取适当的措施，及时修订开发计划，使项目处于可控制之中。

● 阶段跟踪

阶段跟踪主要跟踪每个阶段的完成情况，对应到比较详细的每个阶段的计划。阶段计划是在总体计划的基础上，由阶段负责人制订的，比较详细。包括估计该阶段的工作产品的规模、工作量、资源、成本、风险、进度、任务的划分，以及相对简单的任务计划等等。所以在这个层次上分别跟踪每个阶段的软件产品的规模、成本、资源、风险、进度等。并将实际值与项目计划中的估计值进行比较，如果该偏差值超出定义的范围，则给项目经理、阶段负责人、质量保证人员等发出警告。提醒其采取适当的措施，及时修订计划，使后续工作能顺利进行。

● 任务跟踪

任务跟踪主要跟踪每个任务的完成情况，对应到每个任务的计划。随着阶段计划的实施，任务的划分越来越细，而角色职责的明确，使得每个任务的负责人可以对自己的任务制定出详细的计划。因此对每个任务进行跟踪是很方便的。跟踪每个任务的具体实施过程，如任务的实际起始时间、进度、完成率等等，并与计划值进行比较，计算出偏差值。如果该偏差值超出定义的范围，则给阶段负责人、任务负责人、质量保证人员等发出警告。提醒其采取适当的措施，及时修订计划，保证开发过程的可控性。

从上所述可以看出，跟踪的大、中、小层次之间存在一种包含关系。总体计划中包含简单的阶段计划，阶段计划中包含粗略的任务计划。而跟踪可以实现对这三个计划的及时修订。因为计划的实施，总是从最小的地方开始，所以通过对每个任务的跟踪，及时调整任务计划，随着偏差的累加，必然会引起阶段负责人对阶段计划的重新审查；通过对阶段任务的跟踪，及时调整阶段计划，随着偏差的累积，也必然会引起项目经理对总体计划的修订。因此，可以看出，这三个层次的跟踪是一个环环相扣的过程，互相改进的过程。随着跟踪过程的实施，肯定会对项目的顺利实施提供很大的帮助，成为项目管理的好助手。

总结

计划是跟踪的基础。详细的计划可以提高跟踪的准确性，提高跟踪的效率和效果。粗糙的计划则会加大跟踪的工作量，并降低跟踪的效果。因此，制定一个好的、详细的项目开发计划是跟踪的关键。项目跟踪主要针对计划，是为了了解项目的实际进展情况而采取的活动。如了解成员的工作完成情况，了解整个项目计划的完成情况等等。跟踪主要是为了及时了解项目中的问题，并及时解决，不使问题淤积而酿成严重后果。因此在项目管理中，应该将项目计划和跟踪紧密地结合起来，发挥最大的功效。本文将项目计划和项目跟踪分别从三个层

次来设计，是一种切实可行的实现方式。

参 考 文 献

- [1] Software Engineering Institute. *Transitioning Your Organization from Software CMM Version 1.1 to CMMI-SW Version 1.0*. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University
- [2] CMMI product Team, *Capability Maturity Model Integration (CMMISM) version 1.2*. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University
- [3] CMMI product Team, *CMMI for Acquisition, Version 1.2 CMMI-ACQ, V1.2*. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University

作者简介

刘卫宏：1970 年生，女，硕士，陕西省宝鸡市，副教授，主要研究方向：软件过程管理与度量。E-mail: lwh_90@163.com

基于企业营销渠道管理的信息服务

宋 瑾

(工业和信息化部电子科学技术情报研究所)

摘 要: 随着市场竞争的加剧, 营销渠道管理已经成为企业占领市场、维持市场竞争地位的关键环节。在渠道管理的过程中, 渠道网络的整合和优化、渠道策略的实施和控制以及营销渠道的评估和反馈都必须依赖于信息服务。面向营销渠道管理的信息服务需要明确服务内容和策略, 帮助企业有效实施渠道管理、不断优化营销渠道。

关键词: 信息服务; 渠道管理; 市场竞争

The Information Services for Marketing Channel Management

SONG Jin

Abstract: As marketing competition intensifies, the management of marketing channel has become a crucial element by which corporate occupy the market and maintain market competitive position. In the process of channel management, integration and optimization of the channel network, implementation and control of the channel strategy, assessment and feedback of marketing channels have to rely on information services. The information services for marketing channel management need to define the service content and service strategy in order to help corporate to implement effectively channel management, and constantly optimize marketing channels.

Keywords: information services, channel management, marketing competition

1 引言

随着经济全球化, 全球一体化的不断深入, 企业面临的市场竞争越来越激烈。企业的营销策略必须要适应企业内外部营销环境的不断变化。渠道管理是营销管理的重要组成部分, 也是企业商业模式和核心竞争能力的构成要素。企业营销渠道管理的过程是渠道不断整合优化的过程, 而推动和完成这一过程的关键要素就是信息, 因为全面掌握渠道生存和发展的各种信息是营销渠道管理的前提条件。信息服务提供商必须以一定的策略将这些信息提供给渠道管理和决策者, 帮助其优化渠道策略, 创造经济利润。从这个意义上说, 成功的营销渠道管理必须依赖高质量的信息服务。

目前,信息服务已经渗入到政治、经济、社会文化生活的各个方面。对于企业而言,获取和利用信息的能力早已成为企业的核心能力。但实践证明,若信息服务不能与特定的服务对象、特定的事件和特定的环境相结合,只能造成信息的无效和泛滥,反而影响企业的决策。对于企业的渠道管理而言,所需的信息服务必须与渠道管理的特点、目标以及所处的环境紧密结合。当前信息服务在企业营销渠道管理中的应用已经不断地拓展和深化,从基础信息和基础信息技术的提供到渠道商业流程的再造,可以说渠道变革已经成为信息服务的重要实践活动。

2 基于信息服务的营销渠道管理

按照科特勒的定义,营销渠道是指某种货物或劳务从生产者向消费者移动时取得这种货物或劳务的所有权或帮助转移其所有权的所有企业和个人。典型的营销渠道由生产商、批发商、代理商、零售商等实体共同组成。货物或劳务通过营销渠道的各个实体预先形成的契约关系实现流通,并将这些实体以实物转移为纽带链接起来,分享在实物转移中产生的流通价值。通常货物或劳务在营销渠道中的流通是单向、静止的。

渠道管理是指企业营销渠道的整合优化、协调管理与评估反馈的过程。而信息服务贯彻渠道管理过程的始终,发挥着及其重要的作用。甚至可以认为,信息技术和信息服务在营销渠道管理中的广泛应用已经深刻地改变了传统渠道管理的基本形态、管理模式和管理理念。

2.1 信息服务在营销渠道整合优化中的应用

营销渠道的整合优化是指企业根据产品特性、消费者特性、市场竞争特性和企业特性建立健全最适合企业营销战略的营销渠道以及渠道策略。根据上述特征,企业可以选择不同的渠道类型,包括营销渠道的层次和宽度以及对应的渠道策略。因为渠道整合和优化的目标是应对企业内部资源和企业外部营销环境的变化,其实质是在分析外部和内部环境的基础上,结合企业的营销战略和市场目标完善营销网络体系,以最大限度地提升营销渠道的营销效率以及营销渠道的服务产出水平。因此面向营销渠道整合优化过程的信息服务必须与企业进行充分的沟通交流,掌握企业的经营理念,明确企业的细分市场、市场定位、市场目标,保证信息服务的内容和质量与企业的经营战略保持一致性和稳定性,引导服务对象及时准确地掌握企业营销渠道面对的外部机会和威胁,企业拥有的自身优势和劣势等关键信息,以便在渠道优化的过程中,灵活应对各种问题和障碍。

信息服务在营销渠道整合优化过程的主要应用是为企业提供渠道整合优化的系统解决方案,满足解决方案所需的数据信息需求以及信息技术平台的建设。目前,信息服务在渠道整合优化中的主要应用包括渠道网络规划信息系统、网点和经销商选拔系统、商圈研究模型和标杆研究模型等,这些系统通过对信息的搜集、整合、分析为企业提供全方位的渠道解决方案。

2.2 信息服务在渠道协调管理及缓和冲突中的应用

渠道的协调管理是指在企业营销体系的架构下,通过各种经济手段和奖惩机制引导各渠道主体的经营方向、经营方式和营销策略与企业的营销战略相互适应。由于渠道主体均是追求经济利益最大化的经济主体,因此,企业不可能制定出一套制度使各主体之间的利益分配

完全一致。利益冲突则显著体现在同一品牌的渠道分配冲突、同一渠道的品牌冲突以及营销渠道的上下游冲突。因此,渠道的运行需要监督、管理和协调。

尽管如此,渠道冲突仍然不可避免,因为客观上,渠道冲突的另外一个主要原因来源于信息孤岛的存在。比如流通渠道由于各层级中间商的信息不对称,产生货物短缺、货物积压、供需剧烈波动等现象。面向渠道协调管理和缓和渠道冲突的信息服务,必须着力于解决这一问题,消除信息孤岛。目前信息服务在渠道协调管理及缓和冲突中的应用主要有三个方面:

(1) 提供覆盖渠道网络的供应链信息系统。将企业的 ERP 系统与中间商的信息系统相互链接,减少营销渠道的流通成本和交易费用,减少渠道堵塞、供需波动、货物积压的成本;

(2) 提供中间商执行渠道策略的信息和情报。帮助企业有效了解经销商的销售服务质量、经销商对商务政策的执行情况、竞争企业及标杆经销商的销售服务状况,比如“神秘客户系统”等。近年来,很多企业建立了渠道终端信息系统以实现企业渠道管理的在线化、实时化、动态化。通过掌握中间商的上述信息制定相应的策略以缓和渠道冲突;

(3) 提供优化渠道服务标准的策略。结合企业的关键性服务指标,结合市场特点并基于动态调整的原则,提供优化渠道网点服务标准体系的相关信息,以此作为实施渠道网点分级管理、能力提升的依据,从而有力地推动企业营销目标的达成。

2.3 信息服务在营销渠道评估反馈中应用

营销渠道的评估反馈是指企业可以即时掌握中间商以及终端客户对于产品、服务、渠道运行效率的评价结果,即时掌握产品以及服务的改进和需求状况,从而为企业进一步优化供应链体系和营销渠道体系提供客观真实的即时信息。

面向渠道评估反馈的信息服务,其重要目的在于将传统营销渠道单向静止的信息流变成双向交互式、充分流动和共享的信息流,最大限度降低流通成本,提高流通的效率。在渠道管理的过程中,及时有效的信息反馈和信息分析是非常重要的。外部环境的监测可以帮助企业时刻了解它的渠道现状,更好的把握未来市场演变的趋势。分析实际业绩与预期目标的差距,能够了解渠道管理过程中的工作偏差,发现渠道策略制定的失误或者环境带来的影响,提醒企业及时反思营销渠道策略的正确性和适应性。

信息服务在渠道评估反馈中的应用主要体现在客户满意度信息的搜集、分析和研究。客户满意度研究可以帮助企业了解其渠道网点的服务现状及变化、分析消费者的消费特点及期望、搜集顾客不满的环节及原因,以不断提升销售或服务水平。根据企业的个性化需要,信息服务提供商还可以量身定制开发网络实时监控信息系统。企业及其经销商可在线实时查询、统计和分析客户对销售和售后服务的满意度,方便及时发现问题、改进问题。

3 面向渠道管理的信息服务策略

信息服务的本质是服务者以特定的策略和内容帮助服务对象解决问题的社会行为,包括服务对象、服务内容、服务策略和服务者四个要素。

金燕的研究(情报科学 2008)认为:在广泛的实践中,信息服务强调“用户导向”的服务对象,“专业”的服务者,“技术支撑”的服务策略,“依存”、“交互”的服务条件,“动态”、“适时”的服务过程,“模糊”的服务形式,“社会”、“独立”的服务性质,“针对”的服务功能,

“实效”的服务结果。信息服务策略是指信息服务活动中必要又充分的方式和手段的组合和运用。营销渠道管理的方式是综合的，所涉及的维度也是多种多样的。信息服务的主要策略包括服务过程电子化、解决方案流程化以及信息评估多元化。具体包括如下几个方面。

3.1 服务过程电子化

过程电子化表现在对客户管理信息系统（CRM）及合作伙伴关系管理系统（PRM）及电子商务系统（EC）等应用系统的使用上。现代网络技术的发展，给企业利用网络缩短时空的限制参与竞争创造了前所未有的机遇与挑战，对于构成其价值链的市场前端的渠道体系电子化则成为渠道管理的必然发展趋势。目前很多企业的渠道体系已经建立了电子商务系统（EC），从其使用的效果来看，已经为渠道体系核心能力的提高发挥了不可忽视的作用；

3.2 解决方案流程化

随着 IT 技术的成熟，信息服务提供商关注的重点已经从技术转向了用户。单纯提供基于信息技术的软硬件产品已经远远满足不了服务对象的需求。信息服务者只有设计出既充分考虑用户目前的应用基础又符合未来应用趋势的整体解决方案，才能真正吸引用户。IBM 的转型表明信息服务上不但要提供技术平台等硬件，更要在挖掘各行业深层应用的基础上，提供整体解决方案甚至是商业咨询。从技术和 IT 专业服务逐步过渡到为客户提供商业流程的咨询。这样才能切入到与客户未来发展相关的核心领域。

3.3 信息评估多元化

为了保证渠道管理的效果，企业对渠道实体和终端用户的评估将成为一个重要内容。评估标准丰富化、纬度多元化将是企业渠道管理发展的另外一个重要方向。渠道评估依赖于信息服务产生的数据以及数据分析处理技术。渠道管理要求信息系统能够准确全面地反映渠道的真实绩效水平和发展潜力。

4 结束语

科斯认为，企业规模的临界点是企业扩张需要的管理协调成本等于市场交易成本。随着企业规模的不断扩大，企业营销渠道网络将会越来越复杂，这意味着营销渠道运作耗费的交易成本也可能越来越大，而减少渠道主体之间交易成本的基本手段就是渠道信息化。基本上渠道信息化降低了渠道运行的交易成本，这等同于企业有了做大做强基础。这就是企业管理者在营销渠道变革中越来越倚重信息服务的重要原因。另外，经济全球化、全球一体化也将企业置身于一个开放的系统，企业营销渠道决策者为深入分析其面临的不断变化的内外部环境，对信息产生了更具体更专业的需求，这种需求推动了信息服务业的高速发展。同时营销渠道管理本质上也是信息服务活动的一个重要的社会实践。成功的营销渠道管理在营销渠道建立、渠道策略实施和渠道控制的整个过程中，都完全依赖于高质量的信息服务。因此，作为面向渠道管理的信息服务提供商应该从信息服务的角度深入分析企业的营销渠道管理，明确信息服务在渠道管理过程中的服务对象、服务内容和策略的特点和实质，帮助企业

更好地进行渠道的整合和优化、渠道策略的制定和实施以及渠道的控制和反馈，保证企业在激烈的市场竞争中创造出更多的经济利益。营销渠道管理是信息服务的实践，信息服务是营销渠道管理的核心。

参 考 文 献

[1] 金燕.面向战略管理的信息服务.情报科学.2008:6
[2] 马丁法伊.战略企业管理系统—21 世纪的工具[M].北京:人民出版社,2004:6
[3] 葛存山.论电子商务环境下营销渠道的革新.电子商务.2003:3

作者简介

宋瑾，女，天津，工业和信息化部电子科学技术情报研究所 研究领域：行业信息管理；信息技术与服务

基于DEA方法的商业银行信息化效率评价

王江涛¹ 邱月²

(1. 首都师范大学 政法学院, 北京 100037;
2. 首都经贸大学, 信息学院, 北京 100026)

摘 要: 商业银行信息化评价系统的建立对商业银行信息化发展具有重要意义, 本文依据商业银行信息化评价指标体系, 建立了一种针对商业银行信息化效率评价的 DEA 模型, 最后就模型进行了实例分析。

关键词: 商业银行信息化; DEA 方法; 效率评价

Efficiency Evaluation of informatization for Commercial Banks based on DEA Models

WANG Jiang-tao¹ QIU Yue²

(1. School of Political Science and Law, Capital Normal University, Beijing 100037, China;
2. School of Information Technology, Capital University of Economics and Business, Beijing 100026, China)

Abstract: The construction of assessment system on informatization for commercial banks is of great significance for the development of informatization for commercial banks. So, The DEA model of efficiency evaluation of informatization for commercial banks is established in this paper according to the assessment index system of informatization for commercial banks. The model has been finally experimented with practical instances.

Keywords: informatization for commercial banks; the DEA model; efficiency evaluation.

信息化已经成为中国经济与社会发展最重要的推动力, 大力推动全社会的信息化, 以信息化带动工业化, 这一战略已经取得了可喜的成果。商业银行对信息化的价值也寄予很高的期望, 国家提出了国民经济信息化的发展战略, 商业银行信息化作为国民经济信息化的重要组成部分, 必须根据自身业务发展和信息化建设的需要, 通过多角度和多目标评价分析, 评估信息化实施的状态、效果、效率、效益, 发现存在的潜在问题, 纠正管理方面偏差。有利

王江涛 (1972—), 1972 年生, 男, 博士后, 新疆人, 工程师, 联系电话 15899196688, 研究方向为科学技术哲学, 电子邮件 wangjt163@163.com; 邱月 (1980—), 女, 黑龙江人, 汉族, 博士, 主要从事信息技术研究。

于商业银行科学地推进战略目标，加强内控与管理，防范 IT 操作性风险，完善信息化服务水平，提高信息化的投资收益。通过这种微观层面的评价，促使商业银行正确认识 IT 的作用，理性进行信息化投资^[1]。本文结合商业银行信息化特点，在综合分析国内外现有评价方法的基础上，依据商业银行信息化评价指标体系，建立了 DEA 评价模型，并就模型举例分析。

1 DEA模型

数据包络分析是美国著名运筹学家 A. Charnes, W. W. Cooper 和 E. Rhodes 等学者在“相对效率评价”概念基础上发展起来的一种新的系统分析方法，它主要采用数学规划方法，从投入与产出的角度来评价决策单元（Decision Making Units, DMU）的相对有效性。这种相对有效性是指被评价的决策单元在输入一定的资源投入后，是否规模和技术都发挥到了最佳水平，从而得到了应有的产出^[3]。它以相对效率概念为基础，按照多指标投入和多指标产出，对同类经济系统的相对有效性进行评价的一种新方法。与其他评价方法不同，DEA 方法最大的优势在于模型中选取指标的量纲可以不同，不同的度量单位对 DEA 模型的结果并无影响，DEA 方法其模型的构成形式和求解过程都是应用线性规划的理论^[8]。目前,DEA 方法广泛应用在许多领域，该模型主要是用来研究多指标输入、多指标输出的决策单元同时为“规模有效”与“技术有效”的一种十分有效的方法^{[2][4][5]}。

1) 基于负产出的 DEA 模型

在评价商业银行信息化效率时，通常是输出越大越好，但是信息安全事件却刚好相反，信息安全事件作为决策单元的输出，从商业银行信息化的角度，它是一种不希望的输出，商业银行应尽量减少这种输出，如果要在信息化评价中应用 DEA 模型，关键是解决信息安全事件作为一种无效的输出如何引入模型中。实际上，不管商业银行采用何种信息安全技术，只可能不断减少信息安全事件发生的次数、时间、终端影响数量，而不能完全避免。信息安全事件作为一种负产出，其产生就如同生产中投入生产要素就能生产出产品一样，是不可避免的。

在 DEA 模型效率评价的 C²R 模型中引入负产出输出项，并假设每个决策单元有 t 种负产出输出项，得到下面的模型^[9]：

$$(PE) \left\{ \begin{array}{l} \min \left[\theta - \varepsilon (E_m S^- + E_t P^- + E_s S^+) \right] \\ \text{S.T.} \quad \sum_{j=1}^n \lambda_j X_j + S^- = \theta X_{j0} \\ \sum_{j=1}^n \lambda_j P_j + P^- = \theta P_{j0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_{j0} \\ \lambda_j \geq 0, j=1, 2, L, n; S^+ \geq 0, S^- \geq 0 \end{array} \right.$$

$$(PE') \left\{ \begin{array}{l} \max [a + \varepsilon(E_m S^- + E_t P^- + E_s S^+)] \\ \text{S.T. } \sum_{j=1}^n \lambda_j X_j + S^- = X_{j0} \\ \sum_{j=1}^n \lambda_j P_j + P^- = P_{j0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = a Y_{j0} \\ \lambda_j \geq 0, j=1, 2, L, n; S^+ \geq 0, S^- \geq 0 \end{array} \right.$$

式中, P_j 为 DMU_j 产出负产出输出向量, $P_j = (p_{1j}, p_{2j}, L, p_{ij})^T$, $P^- = (p_1^-, p_2^-, L, p_t^-)^T$ 为松弛向量。

在 DEA 模型效率评价的 C^2GS^2 模型中引入负产出输出项, 并假设每个决策单元有 t 种负产出输出项, 得到下面的模型:

$$(PG) \left\{ \begin{array}{l} \min [\theta - \varepsilon(E_m S^- + E_t P^- + E_s S^+)] \\ \text{S.T. } \sum_{j=1}^n \lambda_j X_j + S^- = \theta X_{j0} \\ \sum_{j=1}^n \lambda_j P_j + P^- = \theta P_{j0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_{j0} \\ \sum_{j=1}^n \lambda_j = 1 \\ \lambda_j \geq 0, j=1, 2, L, n; S^+ \geq 0, S^- \geq 0 \end{array} \right.$$

$$(PG') \left\{ \begin{array}{l} \max [a + \varepsilon(E_m S^- + E_t P^- + E_s S^+)] \\ \text{S.T. } \sum_{j=1}^n \lambda_j X_j + S^- = X_{j0} \\ \sum_{j=1}^n \lambda_j P_j + P^- = P_{j0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = a Y_{j0} \\ \sum_{j=1}^n \lambda_j = 1 \\ \lambda_j \geq 0, j=1, 2, L, n; S^+ \geq 0, S^- \geq 0 \end{array} \right.$$

式中, P_j 为 DMU_j 产出负产出输出向量, $P_j = (p_{1j}, p_{2j}, L, p_{ij})^T$, $P^- = (p_1^-, p_2^-, L, p_t^-)^T$ 为松弛向量。

对于多种信息化生产方案, 先用模型 (PE) 判断是否在技术和规模上有效, 若方案是技术和规模上有效, 则此方案是符合要求的, 然后再用模型 (PG) 判断是否技术有效, 以用来评价某方案在现有资金、技术、管理的基础上, 在一定产出的条件下, 是否最大限度地利用

自生技术条件来尽可能减少生产要素的投入，尽可能减少负产出的输出，以达到相对技术有效，对于 DEA 非有效或弱有效的技术方案，可进一步调整投入-产出指标，使该方案转为有效方案，这是 DEA 方法对商业银行信息化进行投入优化、加强信息安全、增加科技效率的关键。

第 k 个决策单元的投入和产出方面的改进目标值（优化值）：

投入改进目标值：
$$X_k' = \theta_k^* X_k - S^{-*} = \sum_{j=1}^n \lambda_j^* X_j ;$$

负产出可能的目标改进值：
$$P_k' = \theta_k^* P_k - P^{-*} = \sum_{j=1}^n \lambda_j^* P_j ;$$

产出改进目标值：
$$Y_k' = Y_k + S^{+*} = \sum_{j=1}^n \lambda_j^* Y_j 。$$

利用 C^2R 模型还可以判断各方案的规模收益状况：

$$\sum_{j=1}^n \lambda_j^*$$

① 若 $\frac{\sum_{j=1}^n \lambda_j^*}{\theta_k^*} = 1$ ，则表示第 k 个方案在目前的产出条件下，投入规模和负产出都处于最适水平，规模收益良好；

$$\sum_{j=1}^n \lambda_j^*$$

② 若 $\frac{\sum_{j=1}^n \lambda_j^*}{\theta_k^*} > 1$ ，则表示第 k 个方案规模收益递减，即再增加投入量时，负产出会不断增大，而产出增加的效率不高；

$$\sum_{j=1}^n \lambda_j^*$$

③ 若 $\frac{\sum_{j=1}^n \lambda_j^*}{\theta_k^*} < 1$ ，则表示第 k 个方案规模收益递增，即再增加投入量时，虽然会增加一定的负产出，但可以使产出有较大效率的增加。

2 DEA 模型在商业银行信息化效率评价中的应用

1) DEA 模型指标体系的确定

商业银行信息化本身就是一项多输入和多输出的复杂活动，输入和输出的各变量指标，因时间、地点和具体对象的不同其轻重很难确定，运用 DEA 方恰好可以避免预先确定指标权重的困难，为进行客观评价创造了条件。其次，商业银行信息化过程中的各投入与产出很难明确确定其对应关系，商业银行在某一时期的信息化结果，如利税总额增加，不仅仅是一种活动所造成的可能是由多项活动共同参与才能得到的，并且具体由哪项活动参与，商业银行很难区分，因此，使用 DEA 方法就完全避免了这种区分对应关系的困难，使评价工作具有可操作性。基于商业银行信息化的具体特点和 DEA 方法的这些优点，选择 DEA 方法评价商业银行的信息化，是具有一定的使用价值和科学性的，所选择的方法与所需解决的问题是相符的。

应用 DEA 方法对商业银行信息化建设的相对有效性评价，首先要求被评价的决策单元 DMU 具有同种类型，即个决策单元 DMU 具有相同的投入项指标和产出项指标。其次应该建

立相应的评价指标体系。商业银行信息化过程包括信息的传输环境、网络环境、应用环境和服务环境四个组成部分。信息化建设的投入具体可以分为“硬件”和“软件”两个方面的内容。“硬件”建设的投入如设备购置、网络租赁、系统开发、数据挖掘、人员培训等。“软件”建设的投入有制度建设、安全漏洞的研究与解决以及与商业银行信息化相适应的管理机制、经营模式和业务流程整合的投入等。由此确定商业银行信息化建设投入的资源为：

- ① IT 信息管理人员总数（单位：人）；
- ② IT 网络系统、设备投入费用（单位：万元）；
- ③ IT 系统开发、应用、维护费用（单位：万元）；
- ④ IT 人员的教育培训费用（单位：万元）。

商业银行信息化建设的结果将为商业银行今后的发展带来许多发展的机会，例如商业银行市场份额的扩大，运营效率的提高，利税总额的增加等等。为了反映信息化建设投入的效果，我们在这里选取的产出指标有商业银行信息化建设信息安全事件相对应的负产出指标和商业银行信息化建设相对应的增量指标。商业银行信息化建设产出的指标为：

- ① IT 系统中断的次数（单位:次/年）；
- ② IT 系统中断的时间（单位:小时/年）；
- ③ IT 系统中断影响的终端数量（单位:台/年）；
- ④ 劳动生产率提高值（单位:元/人）；
- ⑤ 利税总额增加值（单位:万元）；
- ⑥ 销售收入增加值（单位:万元）。

下面是某地区的四家商业银行进行信息化建设的投入和产出的情况如表 1 所示。

表 1 四家商业银行进行信息化投入与产出的情况表

投入	DMU ₁	DMU ₂	DMU ₃	DMU ₄
1	43	50	52	65
2	200	260	230	350
3	320	320	380	460
4	120	160	150	160
产出	DMU ₁	DMU ₂	DMU ₃	DMU ₄
1	3	5	2	6
2	40	60	65	50
3	25	50	43	46
4	1200	1350	800	900
5	160	280	180	200
6	1100	1600	1200	1300

2) 四家商业银行信息化效率的 DEA 评价和分析

使用现有的 MATLAB 软件进行 DEA 评价计算^[18]，结果如表 2 所示。

表 2 四家商业银行进行信息化效率值

指标	DMU ₁	DMU ₂	DMU ₃	DMU ₄
C ² R 模型	1.0000	1.0000	1.0000	0.9538
DEA 有效性	有效	有效	有效	无效

决策单元 1, 2, 3 是有效的, 落在 C²R 模型的相对有效前沿面上, 说明这三家商业银行的信息化效率相对这个地区的同一行业的四家企业是有效的, 而决策单元 4 则是相对无效的, 存在投入冗余或产出不足, 说明这家商业银行信息化相对其他三家来说是效率差的。对 DEA 无效的企业的投入和产出进行调整, 根据公式 (1)、(2)、(3) 求其在有效生产前沿面上的“投影” (X'_k, P'_k, Y'_k) 。 (X'_k, P'_k, Y'_k) 作为新的决策单元相对于原来的 n 个决策单元来说, 一定是 DEA 有效的, 它提供将 DMU_{*j*0} 转变为 DEA 有效, 而在输入与输出方必须达到的目的。因此, 可以根据投影数据调整投入要素的配置, 提高商业银行信息化的效率。计算后的投影数据见表 3 所示。

表 3 投影数据 (有效值)

投入	DMU ₁	DMU ₂	DMU ₃	DMU ₄
1	43	50	52	47.9231
2	200	260	230	229.2308
3	320	320	380	344.6154
4	120	160	150	138.4615
产出	DMU ₁	DMU ₂	DMU ₃	DMU ₄
1	3	5	2	3.6923
2	40	60	65	47.6923
3	25	50	43	32.6923
4	1200	1350	800	1285.4
5	160	280	180	190.8
6	1100	1600	1200	1240

可见决策单元 4 经过调整后, 信息管理人员总数由原来的 65 人减少到 48 人, 信息网络系统、设备投入费用由原来的 350 万元减少到 229 万元, 软件系统开发、应用、维护费用由原来的 460 万元减少到 345 万元。信息管理人员的教育培训费用由原来的 160 万元减少到 138.46 万元, 而劳动生产率提高值由原来的 900 元/人增加到 1285 元/人, 利税总额增加值由原来的 200 万元减少到 190 万元。销售收入增加值由原来的 1300 万元减少到 1240 万元。说明决策 4 经过调整后, 相对原来的 4 个决策单元来说是 DEA 有效的, 从决策 4 经过调整后的数据还可以看出, 虽然利税总额增加值减少 10 万, 销售收入增加值减少 60 万, 但是信息网络系统费用、设备投入费用、软件系统开发、维护费、教育培训的成本费用减少量远远大于前者, 这里还不包括经过调整人员减少的成本, 以及劳动生产率的提高。综上所述, 商业银行 4 应从加强科技管理, 提高人员素质, 节约不必要的要素投入, 加强信息安全管理, 实现要素的合理流动和充分利用, 提高商业银行信息化的经济效益。商业银行 4 方案规模收益递

减,再增加投入,安全生产事故增多,但产出效率不高。

3 结论

运用 DEA 这种方法能够以线性规划的手段估计相对的有效生产前沿面,确定各决策单元的 DEA 有效性,从而分析当前各种决策单元的相对效益情况。但是,DEA 方法应用上的假设前提是研究、测算、比较对象具有相同的前沿面,在实际中假设所有的评价单元都是在一个前沿面上是脱离实际的。因此评级单元相对于前沿面的差距除了商业银行科技人员素质和 IT 管理水平等“软技术”外,还可能是由于 IT 基础设施等“硬技术”未达到所研究水平时而产生的差距。本章运用 DEA 方法测算了四家商业银行的技术效率,并对所得的效率值进行了横向和纵向的比较和分析,得出的主要结论是经过股份制改造的国有商业银行信息化效率值基本呈现相对于地区商业银行有效的趋势。

参 考 文 献

- [1] 马莉.商业银行信息化评价指标体系及其评价方法的研究[J].现代制造工程,2005,(3):41-44
- [2] 韩松.几种技术效率测量方法的比较研究[J].中国软科学,2004,(4):147-151
- [3] 魏权龄.评价相对有效性的 DEA 方法-运筹学的新领域[M].北京:中国人民大学出版社,1998
- [4] 李鑫等.DEA 方法在企业经营绩效评价中的应用[J].天津职业大学学报,2006,(6):44-46
- [5] 李思寰.基于 DEA 的企业信息化的效率评价[J].科技和产业,2006,(5):28-31
- [6] 段永瑞.数据包络分析-理论与应用[M].上海:上海科学普及出版社,2006
- [7] 魏权龄.数据包络分析[M].北京:科学出版社,2004
- [8] 刘顺忠.管理运筹学和 MATLAB 软件应用[M].武汉:武汉大学出版社,2007
- [9] 宋新山.MATLAB 在环境科学中的应用[M].北京:化学工业出版社,2008

第 6 部分

信息安全

浅谈信息安全中的加密技术

李左伦 杜春梅 郭鑫

(空军航空大学 航空电子工程系 吉林 长春 130022)

摘要: 本文主要介绍了信息安全中的加密技术, 主要有对称加密技术 (DES) 和非对称加密技术 (公钥 RSA) 等, 重点介绍了加密技术的算法, 简要介绍了加密技术的应用、加密技术的发展趋势。

关键词: 加密技术; 对称加密技术; 非对称加密技术; 信息安全; 算法分析

The Study of Encrpytion Techniques of Information Security

LI Zuo-lun DU Chun-mei GUO Xin

(The Electronic Engineer Department, Air Force Aviation University, Changchun, Jilin, China 130022)

Abstract: this paper introduces encrpytion techniques of the information security, which are mainly symmetrical encrpytion technique ,dissymmetrical encrpytion technique and the encrpytion techniques arithmetic,etc. coinstantaneous briefed the applications and evolutive direction of encrpytion techniques.

Keywords: encrpytion technique, information security, arithmetic analysis

在信息时代, 如何做才能达到使信息系统的机密信息难以被泄漏, 或即使被窃取了也极难识别, 以及即使被识别了也极难篡改, 已经成为信息安全中的热点研究课题。安全解决方案可以分为两大类: 一是以防火墙技术为代表的被动防卫型方案, 另一是以数据加密、用户授权认证为核心的主动开放型方案, 即加密技术。加密技术 (Cryptography) 是一门通过加密算法将明文 (plaintext) 和加密密钥 (encryption key) 转换为密文 (ciphertext) 以保护数据安全的科学。一个优秀的加密算法能够做到, 没有解密密钥, 密文很难还原为明文。

1 加密算法分类

加密技术是对信息进行编码和解码的技术, 编码是把原来可读信息 (又称明文) 译成代码形式 (又称密文), 其逆过程就是解码 (解密)。加密技术的要点是加密算法, 加密算法可以分为对称加密、不对称加密、不可逆加密算法。

1.1 对称加密算法（ DES）

对称加密算法是应用较早的加密算法，技术成熟。在对称加密算法中，数据发信方将明文（原始数据）和加密密钥一起经过特殊加密算法处理后，使其变成复杂的加密密文发送出去。收信方收到密文后，若想解读原文，则需要使用加密用过的密钥及相同算法的逆算法对密文进行解密，才能使其恢复成可读明文（如图 1）。在对称加密算法中，使用的密钥只有一个，发收信双方都使用这个密钥对数据进行加密和解密，这就要求解密方事先须知道加密密钥。对称密钥密码体制从加密模式上可分为序列密码和分组密码两类：序列密码、分组密码。

1. 序列密码一直是作为军事和外交场合使用的主要密码技术之一。主要原理是：通过有限状态随机产生性能优良的伪随机序列，使用该序列加密信息流，得到密文序列。所以，序列密码算法的安全强度完全决定于它所产生的伪随机序列的好坏。产生好的序列密码的主要途径之一是利用移位寄存器产生伪随机序列。目前要求寄存器的阶数大于 100 阶，才能保证必要的安全。序列密码的优点是错误扩展小、速度快、利于同步、安全程度高。

2. 分组密码的工作方式是将明文分成固定长度的组，如 64 比特一组，用同一密钥和算法对每一块加密，输出也是固定长度的密文。对称加密算法的特点是算法公开、计算量小、加密速度快、加密效率高、明文加密后产生的密文大小和明文大小差不多，因此可对文件（大数据量）进行加密。

对称密钥密码体制存在的最主要问题是：由于加/解密双方都要使用相同的密钥，因此在发送、接收数据之前，必须完成密钥的分发。密钥更新的周期加长，给他人破译密钥提供了机会。但其具有加解密速度快和安全强度高的优点，目前被越来越多地应用在军事、外交以及商业等领域。

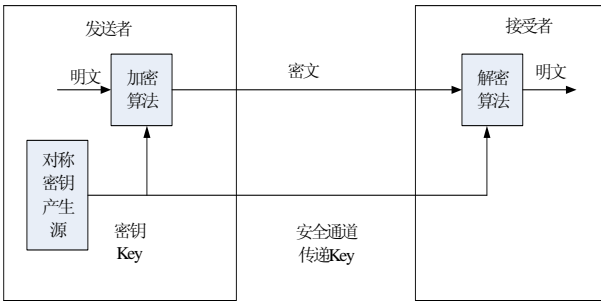


图 1 对称加密原理框图

1.2 不对称加密算法（公钥RSA）

不对称加密算法使用两种完全不同但又是完全匹配的一对钥匙—公钥和私钥。加密明文时采用公钥加密，解密密文时使用私钥才能完成，而且发信方（加密者）知道收信方的公钥，只有收信方（解密者）才是唯一知道自己私钥的人。基本原理是，如果发信方想发送只有收信方才能解读的加密信息，发信方必须首先知道收信方的公钥，并利用收信方的公钥来加密原文；收信方收到加密密文后，使用自己的私钥才能解密密文（如图 2）。由于不对称算法拥有两个密钥，因而特别适用于分布式系统中的数据加密。但不对称加密算法计算量大、加密速度慢、加密效率低、加密后的密文大小比明文大很多，因此该算法只能对少量的数据进行

加密。广泛应用的算法有 RSA 算法和美国国家标准局提出的 RSA。以不对称加密算法为基础的加密技术应用非常广泛。

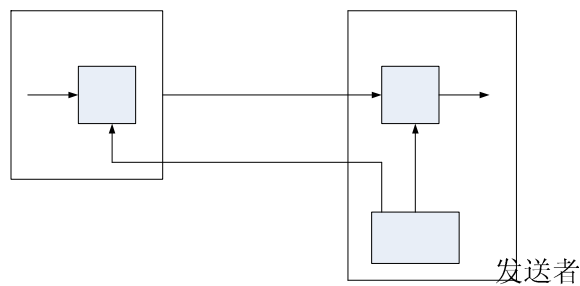


图 2 不对称加密原理框图

2 加密技术的应用

加密算法是加密技术的基础，任何一种成熟的加密技术都是建立多种加密算法组合，或者加密算法和其他应用软件有机结合的基础之上的。

2.1 非否认（Non-repudiation）技术

该技术的核心是不对称加密算法的公钥技术，通过产生一个与用户认证数据有关的数字签名来完成。当用户执行某一交易时，这种签名能够保证用户今后无法否认该交易发生的事实。

2.2 PGP（Pretty Good Privacy）技术

PGP 技术是一个基于不对称加密算法 RSA 公钥体系的邮件加密技术，也是一种操作简单、使用方便、普及程度较高的加密软件。PGP 技术可以对电子邮件加密，防止非授权者阅读信件；还能对电子邮件附加数字签名，使收信人能明确了解发信人的真实身份。

2.3 数字签名（Digital Signature）技术

数字签名技术是不对称加密算法的典型应用。数字签名过程是将数据源发送方使用自己的私钥对数据校验和与数据内容有关的变量进行加密处理，完成对数据的合法“签名”，数据接收方则利用对方的公钥来解读收到的“数字签名”，并将解读结果用于对数据完整性的检验，以确认签名的合法性。在公钥与私钥管理方面，其与 PGP 技术正好相反。在数字签名应用中，发送者的公钥可以很方便地得到，但他的私钥则需要严格保密。

2.4 PKI（Public Key Infrastructure）技术

PKI 技术是一种以不对称加密技术为核心、可以为网络提供安全服务的公钥基础设施。PKI 技术最初主要应用在 Internet 环境中，为复杂的互联网系统提供统一的身份认证、数据加密和完整性保障机制，受到银行、证券、政府等核心应用系统的青睐。

3 加密的未来趋势

尽管非对称密码体制比对称密码体制更为可靠,但计算过于复杂。正是不同体制的加密算法各有所长,所以在今后相当长的一段时期内,各类加密体制将会共同发展。

在对称密码领域,一次一密被认为是最为可靠的机制,但是由于密码体制中的密钥流生成器在算法上未能突破有限循环,故一直未被广泛应用。若找到一个在算法上接近无限循环的密钥流生成器,该体制将会有个质的飞跃。目前混沌学理论的研究给在这一方向产生突破带来了曙光;此外,充满生气的量子密码被认为是一个潜在的发展方向,该理论对于在光纤通信中加强信息安全、对付拥有量子计算能力的破译无疑是一种理想的解决方法。

参 考 文 献

- [1] Andrew Nash 那什,William Duane,Celia Joseph,Derek Brink. 公钥基础设施(PKI):实现和管理电子安全 张玉清,陈建奇,杨波,薛伟译.第一版. 北京:清华大学出版社,2002
- [2] 黄月江, 信息安全与保密, 第二版, 国防工业出版社

作者简介: 李左伦, 男, 杜春梅, 女, 硕士, 空军航空大学航空电子系教员, 主要从事军事通信、导航、雷达、对抗、信息论及信息安全方面研究。

一种改进木材细胞图像分形特征提取方法

任洪娥 高莹 董本志

(东北林业大学, 哈尔滨 150040)

摘要: 木材纹理细胞图像中具有重要的木材纹理信息, 为了对木材树种进行更为深入的研究, 获取图像中的纹理特征参数就显得尤为重要。传统研究方法针对图像整体进行研究, 有着很大的局限性, 于是提出对木材纹理细胞图像不同区域进行分形描述。理论分析和实验结果表明: 该方法能够提取木材纹理分形特征, 并且能够更好地反映木材树种自身不同区域的纹理特性, 这样更有利于对于木材树种的研究, 是木材纹理研究的又一项重要方法。

关键词: 细胞图像; 木材纹理; 分形维数; 特征提取

The Improved Fractal Feature Extraction Algorithm of Wood Cell Image

REN Hong-e GAO Ying DONG Ben-zhi

(Northeast Forestry University, Harbin 150040, China)

Abstract: Wood texture cell image has important wood texture information. In order to deeply study timber species, extracting texture characteristic parameters is especially important. Traditional research method is to examine the overall image. Because it has a lot of limitation, divide the wood texture cell image into different regions, and then extract the fractal dimension of the regions. In terms of the theory analysis and experiments, it is proved that the method can extract the characteristics of wood texture, and it can better reflect the texture features of timber species in different regions of wood texture cell image. The method is good for the study of timber species, and it is it is an important method.

Keywords: cell image; wood texture; fractal dimension; feature extraction

1 引言

纹理是理解图像的一个极其重要的信息源, 所谓纹理可简单地理解为在一个界限了的图像区域中有规律且相互依赖的灰度值的分布, 它具有大小和方向的变化, 从而显示出图像中所反映的不同现象, 它是图像各像素元灰度空间分布局部性质的一种描述, 是描述与识别图像中感兴趣目标的重要依据。而木材的纹理图像正是能够反映不同区域纹理, 具有重要信息

特性的图像资源。任宁、于海鹏等^[1]在对木材纹理进行研究的过程中获取的木材表面纹理信息多是从整体上进行的，没有考虑到木材自身的纹理特点，这样就忽略了局部特征的意义，于是本文从局部区域特征上进行了研究和探讨，为木材树种的纹理研究提供了一种新的方法。

2 分形特征提取

木材纹理是在天然生长过程中形成，具有其自身独特的纹理特征，因此对木材纹理的识别与定量研究就显得尤为重要^[2,3]。木材纹理细胞图像的分形特征是木材独特的纹理特征之一，经过研究表明分形在纹理分析中的应用主要通过求纹理细胞图像的分形维数（Fractal Dimension, FD）^[4]来进行的。在纹理通直情况下，分形维数值能够很好地表征木材纹理的形状、分布密度及其均匀程度，即纹理分布密度越高、纹理宽度越大，分形维数值越大；纹理分布均匀程度越好，分形维数值越小，由此可见木材纹理细胞图像的分形维数可以代表其他很多的纹理特征，具有代表性。

目前，定义分形维数的方法有多种，常见的有相似性维数、容量维数、Hausdorff 维数、信息维数、Lyapunov 维数、谱维数、计盒维数、填充维等。在王克奇的“基于分形理论的木材纹理特征研究^[5]”一文中，就利用分数布朗运动理论求得了木材纹理图像的分形维数值，对木材纹理粗糙程度进行了研究。目前，计盒维数法便于用计算机实现，因而成为应用最广泛的一种分形维数定义方法。

于是，提出采用计盒维数法对木材纹理细胞图像进行分形分析，从而在微观领域上对木材纹理进行了定量化研究，具有快速性、有效等特点。计盒维数法计算过程主要是：

(1) 将 $M \times M$ 灰度图像在 x - y 平面上划分为 $s \times s$ 大小的网格，每个网格处有一列大小为 $s \times s \times s1$ 的盒子，其中选用 $M^{1/3} \leq s \leq M/2$ 且 s 为整数，尺寸因子 $r = s/M$ ，盒子高度 $s1$ 为一个变量；

(2) 检索整幅图像各个网格内的所有像素点，对网格内每一个像素点的灰度值分别进行统计。以第 (i, j) 个网格为例，将跟踪得到该网格内不同种类灰度值计入集合 $Index$ 中，即

$$Index = \{index_{i,j}(1), index_{i,j}(2), K, index_{i,j}(Q)\} \tag{1}$$

可以得到 (i, j) 网格内不同种类灰度值的个数 $n_r(i, j) = Q$ ，进行加和得到总个数值 N_r ；

(3) 最后根据上式计算出的不同 r 值对应的 n_r 值，然后运用以下公式

$$D = \log N_r / \log(1/r) \tag{2}$$

即求得该整幅图像的分形维数值 D 。

应用该计盒维数法计算了包括长白落叶松、红松、鱼鳞云杉、长白松四个树种纹理细胞图像的分形维数值，实验结果表明获得的分形维数值能够在一定程度上表征木材纹理的复杂度等纹理特征，但是仍存在着各个树种之间数值差异较小不利于比较等问题，经过分析表明一部分原因在于该算法主要是针对整幅纹理细胞图像进行分析，这样就造成对于部分区域特性的忽略，这就给树种的分析比较带来了较大的困难。

针对上述问题，提出将木材树种的纹理细胞图像做部分分割，对图像不同部分分别进行分形描述，这样既可以反映不同区域的不同纹理分形特征，使其不被整体所覆盖而造成特征丢失，又可以在进行样本比较的过程中更为准确方便。经过对四个不同树种近千幅纹理细胞图像实验，下面简要列出四个树种几幅图像的不同区域的分形维数值，如表 1 所示。

如图 1 所示为四种已知木材树种纹理细胞图像。

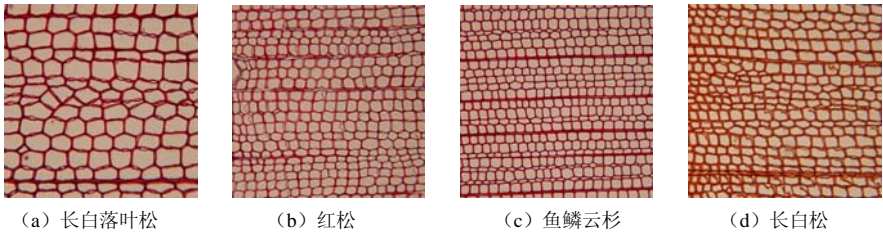


图 1 木材树种纹理细胞图像

表 1 木材树种分形维数数据

	样本	区域 1	区域 2	区域 3	区域 4
长白 落叶松	1	3.1129	3.2199	3.0967	3.2338
	2	3.1011	3.2262	3.0963	3.2397
	3	3.1132	3.2314	3.1033	3.2352
	4	3.1318	3.2026	3.1114	3.2192
红松	1	3.1974	3.1953	3.2115	3.2079
	2	3.1896	3.1963	3.2117	3.2042
	3	3.1972	3.1913	3.2145	3.2033
	4	3.195	3.1979	3.2059	3.2095
鱼鳞 云杉	1	3.2115	3.2121	3.2142	3.2137
	2	3.2123	3.2094	3.2189	3.2109
	3	3.2121	3.2108	3.2156	3.2113
	4	3.2084	3.2128	3.2164	3.2126
落叶松	1	3.2207	3.2248	3.2234	3.2358
	2	3.2206	3.223	3.2269	3.2354
	3	3.2201	3.2202	3.2295	3.2302
	4	3.2116	3.2096	3.2336	3.2312

3 实验结果分析

经过对上面表 1 中的实验数据分析可以看出，在所选取的四种木材样本图像实验数据中，相同树种不同样本图像的各相同划分区域具有数值相近的分形维数值，并且根据不同树种的不同纹理特征这种表现还不尽相同。以长白落叶松为例，其纹理较为不均匀，因此获取的不同区域的纹理细胞图像的分形维数值就根据其不同区域特征有一定差异，如区域一、三主要分布在 3.09 至 3.13 之间，而区域二、四则由于纹理较为不规律取值范围分布在 3.20 至 3.24 之间。而相对于长白落叶松纹理较为规则的鱼鳞云杉四个区域的分形维数值则没有太大的差异，四个区域的取值都普遍分布在 3.20 至 3.22 之间。由此可以看出，这样的划分方法使得木材纹理的区域特性得到了更大限度的体现，更具有科学性与实用价值。

另外，在对检测树种进行分析时，以上方法所获取的分形维数值也有很大的作用。以下

面一幅待检测树种纹理细胞图像为例，如图 2 所示，用上述分形方法获取的不同区域分形维数值如表 2 所示。可以看出虽然单纯的只考虑第四区域取值，该树种可能是红松或是长白落叶松，但是从第一、三区域来看，该检测树种是长白落叶松的可能性就较大，这样就为树种的分析提供了有力信息。在对木材树种的分析过程中，不同区域的不同特征值能够更大程度得到体现，并且不会造成信息丢失。

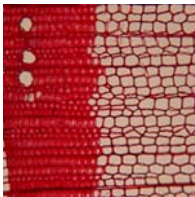


表 2 检测树种分形维数数据

检测样本	区域 1	3.1073
	区域 2	3.2314
	区域 3	3.0983
	区域 4	3.2397

图 2 检测树种纹理细胞图像

4 结论

纹理分析主要包括两个方面的工作：特征提取和纹理分割，其中特征提取是纹理分析的关键所在。分形维是分形的重要特征和度量，它可以把图像的空间信息和灰度信息简单而又有机地结合起来。在本文中将原本的计盒维数法做了部分改进，为了能够更好地体现不同树种的不同区域特性，将原有图像进行了区域划分，实验证明了不同区域分形特征提取的必要性。另外该方法过程方法简单、有效，运用的计算时间较少，效率较高。

因此，采用改进的分形特征提取方法获取木材纹理特征这些规律信息，它不同于以往人们用肉眼进行的观察与主观判别，而是更加具有客观性、科学性，它为木材纹理的分析研究提供了一种思路和方法，具有十分积极的意义。随着人们研究的逐步深入,领域知识的不断完善及广阔的应用前景,相信在不久的将来会有更好更为完善的计算方法出现。

参 考 文 献

[1] 任宁，于海鹏，刘一星，等. 木材纹理的分形特征与计算[J]. 东北林业大学学报，2007，35（2）. 9-11.
[2] 任洪娥，关明山，马岩. 板材的纹理识别初探[J]. 木材加工机械，2004，15（1）. 5-8.
[3] 于海鹏，刘一星，刘镇波. 木材纹理的定量化算法探究[J]. 福建林学院学报，2005，25（2）. 157-162.
[4] Chaudhuri B B, Sarker N. Texture segmentation using fractal dimension[J]. Transactions on Pattern Analysis and Machine Intelligence, 1995, 17（1）. 72-77.
[5] 王克奇，谢永华，陈立君. 基于分形理论的木材纹理特征研究[J]. 林业机械与木工设备，2005，33(7). 19-20.

作者简介

任洪娥（1962-），女，博士，吉林白山人，教授，主要研究方向为图像处理与模式识别、人工智能与智能控制、信息安全；
高莹（1984），女，硕士研究生，主要研究方向为模式识别与智能控制；
董本志（1975-），男，博士生，副教授，主要研究方向为智能控制与计算机仿真。

一种基于TPCM可信移动存储设备的发行与认证

阮富生 王冠 刘智君 王博

(北京工业大学 计算机学院 计算机系, 北京 100124)

摘要: 随着移动存储设备的广泛运用, 由移动存储设备引发的信息泄露的安全问题越来越受到关注, 而将移动存储设备纳入可信体系是解决泄密的一种新的思路和可行的解决办法, 可信移动存储设备的发行和认证是该方法的关键点, 本文提出的基于 TPCM 可信移动存储设备发行与认证具体方法提供了一种解决该关键点的可行方案。

关键词: 可信移动存储设备; 可信计算平台; TPCM; 发行; 认证

A Release and Authentication of Trusted Portable Storage Device Based on TPCM

RUAN Fus-heng WANG Guan LIU Zhi-jun WANG Bo

(Dept. of Computer, Beijing University of Technology, Beijing 100124, China)

Abstract: Along with the widespread of portable storage device, the security of information divulgence initiated by the portable storage device receives more and more attention. Putting portable storage device into trusted system is a new idea and feasible solutions to solve information divulgence, the release and authentication of portable storage device is the key point. This paper provides a solution to the key point.

Key words: trusted portable storage device; trusted computer platform; TPCM; release; authentication

1 引言

移动存储设备作为一种方便快捷的数据存储、传递工具, 在党政机关、军队、科研院所、企事业等单位的核心涉密部门被大量使用^[1], 越来越多的敏感信息、秘密数据和档案资料被存储在移动存储设备里。然而, 各部门对于移动存储设备的管理不是非常规范, 缺乏严格的保密管理措施以及人员使用涉密数据薄弱的安全意识, 导致涉密数据遭遇到了前所未有的挑战。这其中暴露了以下三个主要的安全问题^[2]。

(1) 数据机密性无法得到保证。对于携带处理涉密信息的移动存储设备外出时, 文件在传输中可能被非授权截取, 或在使用过程中通过移动存储设备任意拷贝导致机密数据泄露。

(2) 破坏文件的完整性。内部工作人员主动或被动地拷贝数据, 无意的编辑修改致使机

密数据的完整性、安全性受到很大的威胁。

(3) 恶意代码的威胁。存有机密数据的移动存储设备存在接入网络的可能性，而一旦接入互联网，移动存储设备的机密信息就有可能被他人非法获取。比较典型的是“轮渡”木马。

使用可信的存储设备很好地解决了信息泄密的问题，但是可信存储设备使用存在生命周期，其中发行和认证是两个核心环节，而在该领域研究还仅限于理论探讨，并没有给出具体的实施方案。本文设计的发行和认证方法具有一定的可行性和实用性，具体阐述了从出厂设置，发行注册到认证使用的全过程。

2 可信认证体系

1) 可信计算平台

可信计算平台是一种构建在计算系统中，通过 TPCM^[3]对计算系统实施保护和管理，用于实现可信计算功能的支撑系统。可信计算平台可分为可信平台控制模块，可信移动设备驱动程序和存储介质。

① 可信平台控制模块

可信平台控制模块是一种集成在可信计算平台中，用于建立和保障信任源点的硬件核心模块，为可信计算提供完整性度量、安全存储、可信报告以及密码服务等功能。它与国际可信联盟组织 TCG^[4]提出的 TPM^{[5][6]}处于对等地位，其为移动存储设备存储 TPD ID 列表、根密钥、公钥列表和访问权限列表，在其 PCR 寄存器中存储可信移动存储设备驱动程序的哈希值，以确定驱动程序的合法性。

② 可信移动存储设备驱动程序

可信移动存储设备程序可分为两个部分：可信通信代理驱动程序和标准驱动程序。其中可信通信代理程序是组成结构中的核心原件，负责为可信认证、可信通道等提供核心支持。

③ 存储介质

存储介质指主机硬盘，为可信移动存储提供日志存储等服务。

2) 可信移动存储设备

可信移动存储设备^{[7][8]}（TPD）是指在可信计算平台（TCP）体系下发行的 U 盘或移动硬盘等移动存储设备。在可信计算平台体系下，如果使用移动存储设备进行数据交换，只可以使用经过发行的可信移动存储设备。可信移动存储设备中的可信存储区只允许在同一可信域内同经过验证的可信计算平台进行数据交换，经过发行的可信移动存储设备不允许也无法与非可信计算平台体系的计算机进行数据交换。可信移动存储设备可以按照可信计算平台的要求对数据进行可信封装，保证存储区物理介质中的数据能够抵抗恶意的数据分析。

可信移动存储设备端包含存储区域和可信计算程序接口两部分。其中存储区域被分成了两个部分：可信存储区和公共数据区。可信存储区中存放两类数据：一类是安全属性数据，是为可信存储提供支持的核心数据，对外部是不可见的。另一类是可信数据，是可信域内用户用来交换的数据，受可信存储手段保护。见图 1。

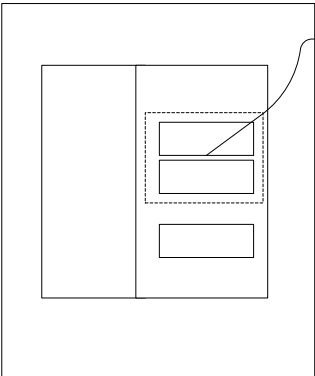


图 1 可信移动存储设备内部结构

3 可信移动存储设备的发行

1) 注册和发行

可信移动存储设备^[9]出厂后设置为初始状态，它具有唯一序列号以区别其他设备。TPD 使用前必须对其进行注册和发行以确定其归属和使用范围，同时建立发行者与设备方相互信任的关系，并为 TPD 提供重要的技术参数。

TPCM 将发行信息写入到 TPD 安全属性存储区中，同时将此信息记录在本地的可信存储区域。设备的注册和发行必须在安全的通信环境中进行，TPCM 须保证 TPD 的注册发行信息被可靠保存，不得随意传播。TPD 需保证自己发行信息存储在安全区域，此区域不易被攻击者非法获取。发行后的 TPD 可以在发行者信任的 TPCM 平台之间交换数据。TPD 的发行者有权指定信任的 TPCM 平台列表，并规定对 TPD 的使用权限。同时 TPD 的发行者有权修改 TPD 的发行信息，甚至销毁其发行信息，使 TPD 重新进入未发行状态。

在 TPD 的发行过程中，有发行和注册两个阶段（见图 2）。

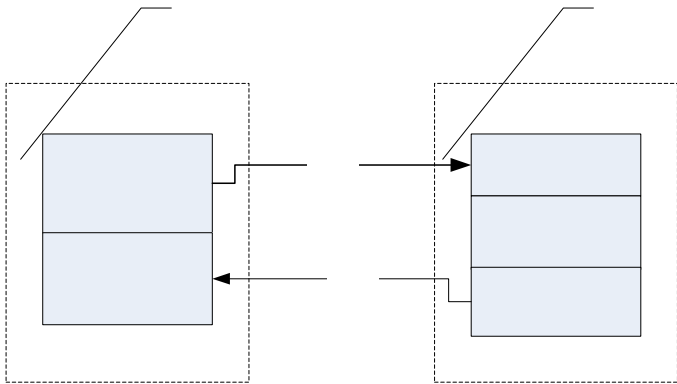


图 2 可信移动存储设备的注册与发行

（1）发行阶段。可信计算平台将自己获得的 TPCM ID 列表写入到 TPD 的信任表中，使 TPD 获悉自己所在的可信域及可信域内的互信主机列表。

（2）注册阶段。TPD 将自己的 TPD ID 写入到可信计算平台中，使可信计算平台对 TPD 建立信任关系，完成注册与发行的过程。

2) 可信域划分与访问控制

可信计算平台系统是由不同区域组成的，每个可信移动存储设备归属于特定的系统，某个可信区域在发行属于自己区域的可信移动存储设备时，可以可信计算平台与可信移动存储设备的相互信任等级以及其他可信域与本可信域可信移动存储设备的相互信任等级，经过发行后的可信移动存储设备内保存有一张信任列表，记录不同区域不同计算平台上的读写控制权限。见图 3。

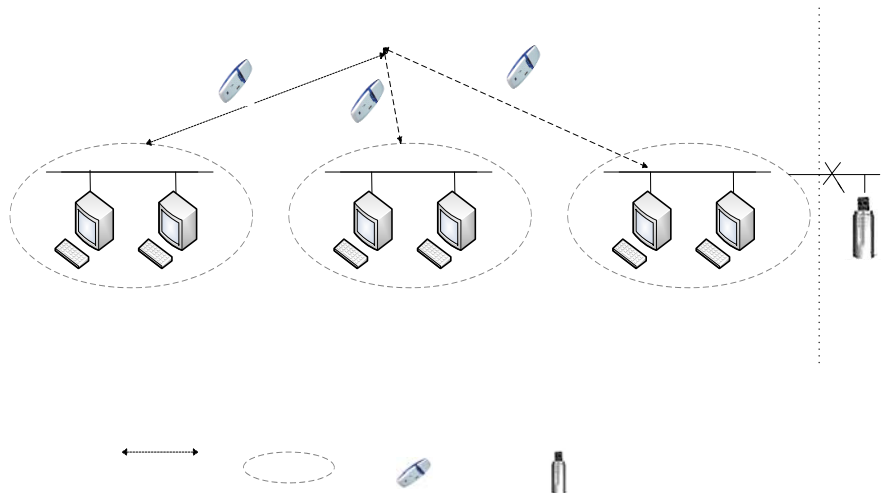


图 3 可信域的划分和访问控制

4 可信移动存储设备与TPCM的双向认证

1) 可信认证的核心模块

要完成可信移动存储设备与 TPCM 可信认证需要涉及到两个核心模块：可信驱动程序和设备固件。

(1) 可信驱动程序端有两个功能模块：双向认证通信代理模块和可信通道的建立与维护模块，双向认证通信代理模块在完成其功能的过程中涉及与 TPCM 的交互过程。

(2) 设备固件端有三个功能模块：双向认证模块、可信通道的建立与维护模块和存储分区与读写权限控制模块。其中，双向认证模块在完成其功能的过程中涉及与可信存储设备的交互过程。

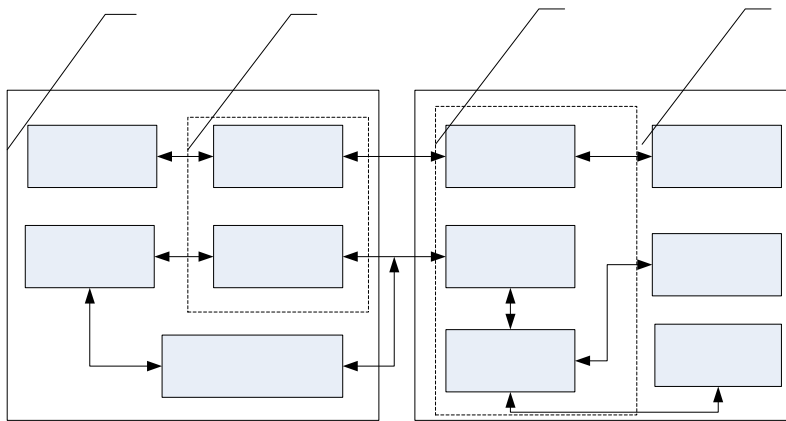


图 4 可信移动存储设备与 TPCM 交互

两个模块通过交互主要完成以下三个功能：

(1) 主机端可信驱动程序双向认证模块与设备固件的双向认证模块负责代理 TPCM 芯片与 TPD 之间的通信数据，帮助其完成双向认证。

(2) 可信通道建立与维护模块负责建立起可靠的数据通信通道，完成硬盘到可信数据存储区的数据传输存储工作。

(3) 存储分区与读写控制权限模块负责管理对可信数据存储区的访问。见图 4。

2) 可信认证方式及过程

可信移动存储设备（TPD）与可信计算平台中的 TPCM 进行相互认证之前，首先必须保证两者之间已经建立好了可信通道,然后需要完成可信认证和可信验证两个过程：

(1) 可信认证阶段：在 TPD 接入阶段，可信驱动程序首先从 TPD 取得 TPD 的 TPD ID，再取出可信计算平台存储在 TPCM 的 TPD ID 列表，在其中查找 TPD 的 TPD ID。若找到，则通过此一阶段认证。若没有找到，则禁止 TPD 运行，结束整个双向认证过程。

(2) 可信验证阶段：在可信认证通过后，可信驱动程序从可信计算平台取得 TPCM 的序列号 TPCM ID，再从 TPD 取出 TPD 信任列表中的 TPCM ID 列表，在其中查找可信计算平台的 TPCM ID。若找到，则通过可信验证。若没有找到，则禁止 TPD 运行，结束整个双向认证过程。流程图见图 5。

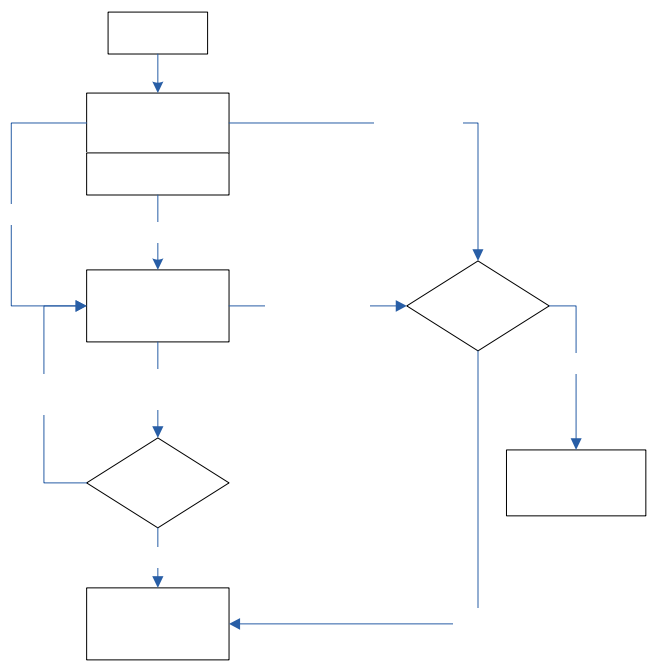


图 5 可信移动存储设备与 TPCM 双向认证流程

3) TPD 工作状态机

认证通过后，TPD 通过查找根据当前用户的信任列表确定当前可信计算平台主机的读写访问权限，产生一个特定的工作状态，假设 T 为其可信数据区，U 为公共数据区：

(1) 一般情况下（默认状态），在可信域内可信主机读写权限是：T:READ&WRITE&~EXECUTE，U: READ& ~EXECUTE；在非可信域：T: NULL，U: READ&WRITE&~EXECUTE；

(2) 在特殊情况下，即非可信域必须从可信域获取数据，则可由 super administrator 在可信域内单次临时开通权限 U: WRITE，使可信域内主机获得临时权限：T: READ&WRITE&~EXECUTE，U: READ&WRITE&~EXECUTE，在单次执行后恢复默认状

态: T: READ&WRITE&~EXECUTE, U: READ& ~EXECUTE。

5 结论

本文设计的基于 TPCM 可信移动存储设备发行与认证的方法,通过对可信移动存储设备的发行和与 TPCM 的双向认证将移动存储设备纳入可信体系,使得可信移动存储设备专属专用,根本上解决了涉密信息被随意拷贝操作进而泄密的难题。为了实现移动存储设备与 TPCM 主机交换数据的完全可信,将来的研究方向和问题主要集中在以下几个方面。

(1) 可信通道。它是指主机、可信平台控制模块、可信存储设备之间通信的信道,是对数据传输过程进行的保护,包括可信通道的建立与撤销和传输数据的保护方式。

(2) 可信度量。对可信存储区、隐藏存储区和普通存储区的完整性鉴别。

(3) 可信存储区域控制。包括对主存储区和隐藏存储区的访问控制、主存储区和隐藏存储区内部的划分、主存储区内部可信存储区和普通存储区的划分、主存储区和隐藏存储区内部存储结构的定义;TPD 在固件层能够实现对自身存储区域的管理和控制,应用层驱动程序能够协助设备完成相应管理工作。

(4) 可信封装。指利用可信平台控制模块(TPCM)和存储设备内部的可信功能模块敏感数据进行封装,包括数据结构组织方法、加密过程和加入鉴别信息过程。

(5) 可信存储日志:包括对可信存储日志审计的数据结构的定义、日志数据存储位置定义、日志分类和日志审计信息生命周期定义。

(6) 可信删除。指对可信存储数据的删除和对普通存储数据销毁。

参 考 文 献

- [1] 移动存储介质管理系统[Z]. 中安网脉(北京)技术股份有限公司.
- [2] 王庆丰,刘功申.一种可信移动存储介质管理系统的设计与实现[J].信息安全与通信保密.2008
- [3] 张兴,沈昌祥.一种新的可信平台控制模块设计方案[J].武汉大学学报.2008,33(10)
- [4] Trusted Computing Group. About TCG[OL]. http://www.trustedcomputinggroup.org/about_tcg
- [5] Trusted Computing Group. Trusted Platform Module[OL]. http://www.trustedcomputinggroup.org/developers/trusted_platform_module
- [6] David Challener, Kent Yoder 等著.可信计算[M].机械工业出版社.2009
- [7] Trusted Computing Group. TCG Storage Workgroup Security Subsystem Class: Optical Specification Version 1.0[S].
http://www.trustedcomputinggroup.org/files/resource_files/8803F27E-1D09-3519-AD274465B5C3E176/Optical_SSC-100.pdf. 2008 September 25
- [8] Trusted Computing Group. TCG Storage Security Subsystem Class: Opal Specification Version 1.0 [S].http://www.trustedcomputinggroup.org/files/resource_files/88023378-1D09-3519-AD740D9CA8DFA342/Opal_SSC_1.0_rev1.0-Final.pdf. January 27, 2009
- [9] Trusted Computing Group. Storage Work Group Optical Storage Summary[OL].
http://www.trustedcomputinggroup.org/resources/storage_work_group_optical_storage_summary. October 2008.

基于P2P文件共享系统的抗攻击信任管理模型

吴旭¹ 郭玉翠² 宫尚宝¹

(1 北京邮电大学 理学院 北京 100876

2 河南理工大学 数学与信息科学学院 焦作 454000)

摘 要: 针对 P2P 文件共享系统提出了一种占用节点内存少、计算复杂度低的信任管理模型。利用二进制向量来同时记录节点的直接信任值和贡献度, 通过 G-Index 方法计算节点的推荐信任值, 由直接信任值和推荐信任值, 通过引入罚函数计算节点的综合信任值。根据节点贡献度及直接信任值计算节点的声望值, 利用该声望值可以有效减少 P2P 文件共享系统中的免费搭车行为。该模型对节点行为变化具有快速的适应能力, 同时对节点恶意推荐具有较强的抵抗性, 和 EigenTrust 算法相比, 计算量大幅降低。

关键词: P2P 系统; 信任管理; 二进制向量; G-Index 方法

1 引言

当今互联网正在飞速发展, 大规模分布式对等网络 P2P (peer-to-peer) 系统得到广泛应用。P2P 系统的出现给人们的资源共享和信息交流带来了极大的方便, 但由于 P2P 系统具有高度的开放性, 匿名性以及动态性, 使得各种病毒容易在 P2P 系统尤其是 P2P 文件共享系统中传播, 在该系统中, 传统的访问控制技术已经难以满足现实安全需要。

为了应对随之产生的各种安全性问题, 各国相关学者已经对 P2P 系统节点间的相互交互设计了各种信任计算模型^[1-4], 来度量和评估系统节点的可信任和可依赖程度, 为相互交互提供依据, 从而建立有效的机制, 保证系统的健康发展。

本文提出一种高效的针对 P2P 系统的信任管理模型, 该模型采用二进制向量^[5]来记录系统节点间的交互历史。二进制向量便于计算、容易实现且具有如下特点: 其最高有效位 MSB (Most Significant Bit) 权值是次高有效位权值的两倍, 左边位数权值总是其相邻右边位权值的两倍。图 1 为二进制向量数值变化图。因为二进制向量具有这个特点, 故可以用二进制数中的 MSB 来记录节点间的最新交互 (即给予最新交互最大的权值), 以此类推, 根据交互时间的不同用二进制向量中不同的位数记录。该方法能够体现新的交互比旧的交互具有更大的权值。在用二进制向量记录节点的直接信任值 (DT, Direct Trustvalue) 和贡献度的基础上, 通过 G-Index 方法计算节点的推荐信任值, 由 DT 和推荐信任值及罚函数值计算节点的综合信任值。根据节点贡献度及 DT 计算节点的声望值, 并以声望值作为节点能否从系统中下载资源的依据。本文的结构安排如下: 第一节给出信任度定义及 DT 表示; 第二节为推荐信任值及综合信任值的计算方法; 第三节给出节点声望值定义, 利用声望值来减少系统中的免费搭车行为; 第四节和 EigenTrust 模型进行了比较, 结果表明本模型较 EigenTrust 模型具有更少的计算量。

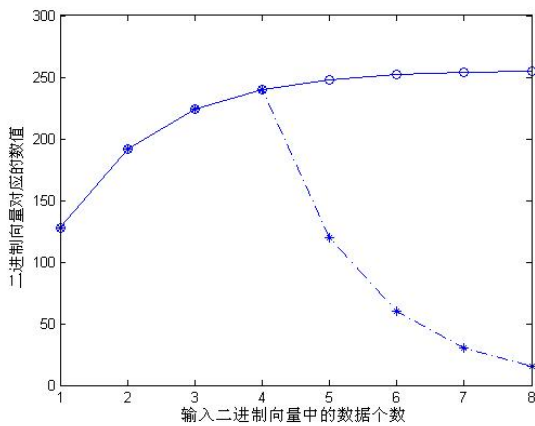


图 1 二进制向量数值变化图

给定一个八位二进制零向量，从 MSB 开始输入数据。每次输入新数据输入到 MSB，其余数据往右依次移动。实线为八位输入全部为 1 的数值变化曲线；虚线为前四次输入为 1 后四次输入为 0 的数值变化曲线。

2 信任度描述

信任度是指网络中一个节点 P 对另一节点 Q 的信赖程度，一般可以由直接信任度和推荐信任度来综合度量。直接信任度主要由 P 和 Q 的直接交互历史来决定，推荐信任度由同节点 Q 有交互历史的其他节点提供给节点 P 的推荐值来决定。直接信任度具有时间衰减性，随着时间的流逝，相互间的交互历史会被“淡忘”，即近的交互比远的交互更具有可信度。直接信任度的表示和计算机制如下。

1) 直接信任表示

在 P2P 系统中，每个节点用一个 n 位二进制数来记录同它交互过的另一节点的交互历史，n 可取 8，16，32 等，n 值越大，则相互交互记录越多，信任值计算越准确，但计算开销也越大。为方便计算本文取 8 位二进制数。值 1 表示成功（或满意）的交互，值 0 表示失败（或不满意）的交互，把新的一次交互值根据交互成功或失败用 1 或 0 记录在二进制数最左边位（即 MSB），把其余位数向右移一位，最后一位丢弃。例如原有记录为 01110011，经过一次成功交互后，则记录更新为 10111001，若交互失败，则记录更新为 00111001。

2) DT 计算

设节点 P 对节点 Q 的原交互记录为 01110011，若新一次交互中 Q 成功提供下载服务，则 P 更新其交互记录为 10111001，为了便于计算 P 对 Q 该时刻的信任值，将其转化为十进制数 185，作为节点 P 对节点 Q 的 DT。同理可得若 Q 提供下载服务失败，则 P 更新其交互记录为 00111001，P 对 Q 的信任值更新为 57。原交互记录 01110011 对应的信任值为 115。每个节点为与其交互过的节点（熟悉节点）建立本地 DT 表，并根据后续交互记录进行更新。本文中 DT 取值范围为 0—255。

表 1 节点 P 本地交互记录更新表

原记录	S / F	更新记录	DT	贡献度
01110011	—	—	115	4
01110011	S	10111001	185	5
01110011	F	00111001	57	4

表 1 中 S 表示交互成功，F 表示交互失败，贡献度定义为二进制向量中 1 的个数，即节点 Q 成功提供下载服务的次数。

3 信任度归集

P2P 系统中节点众多，一般来说与某节点有交互记录的节点只占系统中的小部分，其他大部分节点与该节点没有直接交互记录（称为陌生节点）。若该节点出于某种目的需要和陌生节点进行交互，则需通过其他节点进行推荐，根据推荐结果来最终决定是否交互，以及同哪个节点交互。一般步骤如下。

1) 资源查询

节点 P 在系统中查询它所需要的资源，得到反馈共有 n 个节点拥有该资源并且愿意提供交互，其中 m (m<n) 个节点为陌生节点，其余 n-m 个节点为熟悉节点。若熟悉节点中存在 DT 大于某个阈值（例如 240）的节点，则直接与之交互，否则进入下一步骤：信任度查询。

2) 信任度查询与推荐信任度计算

系统中存在一类节点出于某种目的对某节点给出高于或低于其实际 DT 的推荐值，这种情况称为恶意推荐。为了减少恶意推荐的影响，目前多数信任管理模型主要使用带权推荐方法，即给每个推荐者一个相应的权值（可信度），然后与推荐值作加权平均处理得到最终的推荐信任值，该方法具有一定合理性，但推荐者权值的确定具有很大的主观性，在具体计算中如何给出即符合实际又能够量化的推荐者权值一直是一个难点。为此本文采用 G-Index 方法来计算节点的推荐信任值，G-Index 类似于评价科研人员学术成就的 H-Index，G-Index 既考虑推荐者所给的推荐值，又考虑推荐者的数量，故该方法能够得到一个相对客观的推荐信任值。具体过程如下。

节点 P 向系统中其他节点查询关于节点 Q 的直接信任度，假设从系统中得到 k 个关于 Q 的推荐值，把这 k 个推荐值从高到低降序排列，找到值 L 使得前 L 个推荐值的平均值至少为 L，即 L 满足下式：

$$L \leq \frac{\sum_{i=1}^L DT_i}{L}$$

(1)

记此 L 为系统其他节点关于节点 Q 的推荐信任值。

3) 综合信任度计算

经过信任度查询过程得到 n 个节点的推荐信任度后，利用下式计算 n 个节点中每个节点的综合信任值 T

$$T = \alpha DT + (1 - \alpha)L - f(M)$$

(2)

其中 f 为罚函数，M 为节点 P 对 Q 记录的二进制向量中 0 的个数，即根据以往失败的交互次数对 Q 的信任值进行惩罚。由不同的惩罚力度可以设置罚函数为 βM 或 M^β ($\beta > 0$)。 α

为查询节点的自信系数,取值范围为[0, 1],不同的查询节点其所取的 α 值可以不同,由其自身决定。当被查询节点为陌生节点时 α 取为 0,当被查询节点为熟悉节点时 α 根据查询节点自身的自信程度来取值,若对自身的交互记录越自信,则 α 取值越大。

根据(2)式计算出 n 个被查询节点的综合信任值,在综合信任值排名前 10%的节点中随机选取一个与之交互。这样做的原因如下:对查询节点来说能够下载到可靠的资源,同时对被请求节点来说避免了高信任度节点负载过于集中。

4) DT 更新

若交互对象为熟悉节点,则交互后更新本地 DT 表。若交互对象为陌生节点,则交互后为其建立本地 DT 表,陌生节点转变为熟悉节点。

4 声望 (Reputation) 机制

1) 节点贡献度

在 P2P 系统中,每个节点保存其熟悉节点的 DT 表,DT 大不但说明该节点下载可靠性好,也说明该节点为系统作的贡献大。在 P2P 文件共享系统中,存在大量的免费搭车 (free-riding) 节点,即只从系统中下载资源而不提供资源上传的节点,这类节点的大量存在降低了系统的有效性。为了减少免费搭车行为,鼓励节点分享自己的资源,利用节点本地 DT 表中的信任向量来记录节点的贡献度。由前所述,向量中 1 表示交互成功,即交互节点上传成功,向量中 0 表示交互失败,即交互节点上传失败。显然向量中 1 的个数越多,表示节点为 P2P 系统所作的贡献越大。故可以用该向量中 1 的个数来定义节点的贡献度 (Contribution Value, CV)。如表一所示。

2) 节点声望

现在综合使用节点的 DT 和 CV 来定义节点声望值 (RV, Reputation Value)

$$RV = 1 - \sqrt{\frac{\frac{(DT/255-1)^2}{x^2} + \frac{(CV/8-1)^2}{y^2}}{\frac{1}{x^2} + \frac{1}{y^2}}} \tag{3}$$

式中 x 和 y 为调节参数,通常取为 $\sqrt{2}$ 和 3。RV 是 DT 和 CV 的单调递增函数且值域为 [0, 1]。大量的分析和试验表明^[6],该声望值定义较好地反映了节点声望与 DT 及 CV 的关系。不同节点对某个节点的声望值评价可能不同,为了便于实际应用,可以通过声望查询和归集机制来统一某个节点的 RV^[4]。

对于不同的 P2P 文件共享系统,可以设置不同的 RV 阈值,只有节点的 RV 大于等于该阈值,才允许其在系统中下载资源。对于刚加入系统的新节点来说,其 DT 与 CV 都为 0,根据(3)式知其 RV 为 0,故该节点不能够从系统中下载资源,只能先向系统中其他节点提供资源,建立起该节点的 DT 与 CV,才可能从系统中下载资源,该声望机制有效减少了搭车行为的发生。

5 计算复杂度比较

现采用计算综合信任度中所使用的乘法和除法次数来评估该信任模型的计算复杂度,并

和 EigenTrust 模型进行比较。由式 (1) 和式 (2) 可以计算出 n 个节点的 RT 及综合信任值 T 共需做 $(n+3n)$ 次乘除法, 即该模型计算复杂度为 $O(4n)$, 而在 EigenTrust 模型中, 使用不断向邻居查询的方法来计算最终信任值。例如节点 P 与 Q 没有直接交互记录, 则 P 查询其邻居关于 Q 的交互记录, 计算式为 $T_{PQ} = \sum_{k=1}^n T_{Pk} T_{kQ}$, k 为中间人, 需做 n 次乘法, 若中间人与 Q 仍没有交互记录, 需再次发出查询, 假定查询层数为 m , 则最终总的乘法次数为 n^m , 即 EigenTrust 模型算法复杂度为 $O(n^m)$ 。

6 结束语

本文给出了一种适用于 P2P 文件共享系统中建立节点间信任关系的模型及算法。该算法首先利用二进制向量来刻画节点间的 DT, 该法对节点行为变化具有较好的适应能力; 其次利用 G-Index 方法来获得节点的推荐信任度, 并给出了节点 CV 及 RV 概念。该模型能有效抗击恶意节点的推荐, 同时减少系统中节点的免费搭车行为。同 EigenTrust 模型相比较, 该模型具有更小的计算开销, 在规模较大的 P2P 文件共享系统中有较好的效果。

参考文献

- [1] Blaze M, Feigenbaum J, Ioannidis J. The Role of Trust Management in Distributed Systems Security [C] // Secure Internet Programming: Issues for Mobile and Distributed Objects. Berlin: Springer-Verlag, 1999:185-210
- [2] Kamvar .S. D, Schlosser .M. T., and Garcia-Molina.H The eigentrust algorithm for reputation management in p2p networks. [C] // The Twelfth International World Wide Web Conference, Budapest, Hungary: ACM Press, 2003: 387-1391.
- [3] Xiong L, Liu L. PeerTrust: supporting reputation based trust for P2P electronic communities [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16 (7) : 843-857.
- [4] Zhao H ,Li X . H-Trust: A Group Trust Management System for Peer-to-Peer Desktop Grid [J] Computer Science and Technology .2009 24 (5) : 833-843
- [5] Selcuk A, Uzun Ersin, Pariente M. A reputation-based trust management system for P2P networks [C]// Proc 4th Intel workshop on GP2PC. Chicago, IL: IEEE Computer Society Press, 2004: 251 - 258.
- [6] Zouridaki C,Mark B ,Hejmo M, Thomas R. A Quantitative Trust Establishment Framework for Reliable Data Packet Delivery in MANETs [C] // Proc 3rd ACM Workshop on Security of Ad Hoc and Sensor Networks, Alexandria,VA ,USA,2005
- [7] Hirsch J E. An index to quantify an individual's scientific research output.[C]// Proc. Nat. Acad. Sci., 2005, 102 (46) : 16569-16572.

基金项目: 国家自然科学基金 (项目编号 60973146)。

作者简介

吴旭, 1983 年生, 男, 硕士, 籍贯江西上饶, 助教, 主要研究方向为网络系统中的信任模型与可信计算。

郭玉翠 北京邮电大学 教授 研究方向 信息安全与可信计算。

一种密码算法数据模型的设计与实现

周修义¹ 何新民² 徐莉伟²

(1. 信息工程大学电子技术学院, 河南 郑州 450004; 2. 北京 7223 信箱 10 分箱, 北京 100072)

摘 要: 基于对密码数据专业特征的分析, 本文设计了一种密码数据模型 CipherD, 通过该模型能够清晰地描述密码数据的内在逻辑层次结构。为实现模型 CipherD, 文中提出了三种密码数据的存储模型, 并分别基于以上模型实现了典型的数据抽取操作。通过测试大量的样本数据, 分析比较了三种存储模型的优劣, 进而为全面优化 CipherD 的整体性能提供依据。模型 CipherD 建立了一种密码数据的表示和交换机制, 并进一步为密码算法的标准化描述奠定了基础。

关键词: 密码数据; 数据模型; 存储模型; CipherD

Design and Implementation of a Data Model for Cryptographic Data

ZHOU Xiu-yi¹ HE Xin-min² XU Li-wei²

(1. Institute of Electronic Technology, Information Engineering University, Zhengzhou 50004;
2 P. O. Box 7223 Beijing, Beijing 100072, China)

Abstract: In consideration of the feature of cryptographic data, a new data model of cryptographic data is designed, the logic structure can be reflected clearly with the specification of CipherD. Three memory models of cryptographic data for CipherD model are proposed. And based on the above memory models, the manipulation function of drawing out data is implemented respectively. The advantages of each model are analyzed according to the test results, and can be referred for promoting the performance of CipherD. A kind of mechanism for representing and exchanging cryptographic data is provided by the CipherD model, and the foundation for a further standardized description of cryptographic algorithm is laid.

Keywords: cryptographic data; data model; storage model; CipherD

1 引 言

密码技术是保护信息安全的核心和关键技术, 随着密码技术在硬件和软件系统中的应用日益增多, 对系统中密码模块的测试、验证也变得愈加迫切重要^[1]。然而目前验证密码算法

的工作却非常复杂和烦琐,需要占用大量的时间和精力。密码算法的模拟程序在密码模块测试验证过程中具有桥梁纽带的作用,密码算法数据的表示和相关运算操作则是密码算法实现的基础和核心。

但是,现有的计算机通用程序语言提供的数据类型^[2]不支持密码算法数据的特征,大部分的实现都使用通用数据类型表示密码算法数据,存在随意性大、数据模型不直观、字节序不统一等问题。目前,风格各异的表示方法导致对密码算法的描述比较混乱,间接的表示途径使得对各个编码环节中数据运算的描述冗长、繁琐。

2008年12月,美国Galois公司正式公开发布了针对密码算法的领域专用语言Cryptol^[3],通过Cryptol语言,密码学上的概念能够得到直接形式化的描述,但是涉及政策限制,国内用户仅能获取功能限制的Cryptol工具包,无法进行深入的研究开发。目前,国内仅有文献^[4]针对分组密码算法设计了一种简单的描述语言及其解释器,但是其对于数据的描述支持单一,无法描述密码数据的逻辑层次结构。

为了更加精确地描述密码算法数据,简洁清晰地描述密码算法的各个编码环节,需要对密码算法中的数据进行高层次的抽象描述。本文研究分析了密码数据的专业特征,针对密码数据的逻辑层次结构设计了一种密码数据模型CipherD,并为其设计了三种数据存储模型。基于模型CipherD实现了数据抽取操作,通过测试,分别分析比较了三种存储模型的优劣,从而确定了一种密码数据表示和交换机制,为实现密码算法的标准化描述奠定了基础。

2 密码数据特征

密码算法本质上是数据序列上的数学运算,密码算法中出现的明文、密钥、中间数据、密文等统称为密码数据^[5]。密码数据类型特指密码算法描述中用来表现密码数据的一种数据类型。与通用数据类型相比,密码数据是一种规模伸缩自由、内部操作复杂、外部表现多样的特殊数据,主要有以下几方面明显的特征。

1) 位串特征:密码数据的基本运算,一般建立在二进制位串基础上,且位串长度不固定,有很大的随意性。如:移存器的状态,不同长度的数据(明文或密文)分组等。根据编码的需要,位串长度可以是1位或几位或几百位。

2) 分段特征:处在相同空间或时间的数据整体具有编码上的意义,通常要进行组合使用,这形成了密码数据的分段特征。如,环上移存器各元素数据段的划分,DES算法中数据的左、右部分的划分,一组纯乱码中四个十进制数的划分,多路合成数据中各路数据之间的划分等。编码对数据的分段有着自己的划分和解释,存在随意性。

3) 序列特征:密码算法总是按照某种节拍运行在密码数据的时序和顺序之上,使密码数据具有十分明显的序列特征。如:序列密码算法中各中间环节以及输出环节的密码数据;分组密码DES中16层迭代中使用的16个按照一定顺序排列的子密钥;进入密码算法初始状态时密钥的注入顺序。

密码数据的以上特征使其具有比一般通用的数据类型更加丰富的逻辑层次结构,定长的、不易对其分位进行操作的通用数据类型,难以胜任对密码数据的描述。

3 密码数据模型CipherD的设计

在定义每一个密码数据时，算法设计者实际上在有意或无意间都将其映射到一个二元比特序列，由于映射过程没有统一的标准给密码描述带来了混乱。对于密码数据的描述，实际上是在此二元序列上进行逻辑的划分，从而呈现出不同的数据形式，然而目前通用的描述方法无法表现出这种划分过程。

1) 密码数据模型 CipherD

本文针对密码数据的逻辑层次特征，采用三个数据参量数据高度 H 、数据宽度 W 和数据长度 L 对一个密码数据进行描述， H 、 W 、 L 的取值均为正整数。同时，本文将一个密码数据记作 d ，将其所对应的二元比特序列记作 b 。通过三个参量的描述，一个密码数据的结构如同一个数据立方体，如图 2.1 所示。

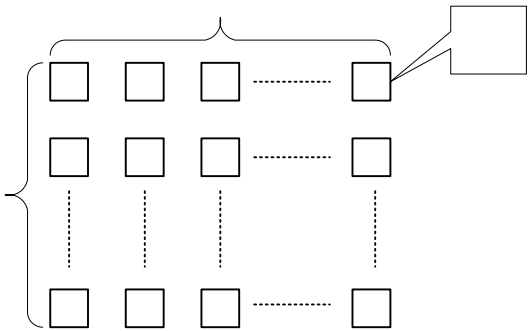


图 2.1 层次结构模型示意图

① 根据密码变换基本运算，将二元比特序列 b 分成相等规模的基本单元，数据高度 H 描述的是基本单元的规模，以二进制位的规模计；

② 根据时序或节拍，将基本单元构成的序列划分为相等规模的节拍，数据宽度 W 描述的是节拍的规模，以基本单元的数目计；

③ 节拍构成的序列是密码数据的最高层表现形式，数据长度 L 描述了序列的规模，以节拍的数目计。

每个密码数据 d 都对应一个二元序列 b ，在将 d 映射到 b 的过程中采用如下规则：

首先将 d 的基本单元按照高位 bit 在左、低位 bit 在右的顺序映射成一个长为 H 的二元序列，记作 u 序列。在一个节拍中，把第 i 个基本单位的二元序列记作 $u[i]$ ，按照以下顺序把一个节拍映射为二元序列，记作 v 序列： $u[W-1], u[W-2], \dots, u[0]$ 。其中， $u[0]$ 是 v 序列的低端， $u[W-1]$ 是 v 序列的高端。

把 d 的第 i 个节拍序列记作 $v[i]$ ，按照以下顺序把 d 映射为 b 序列： $v[L-1], v[L-2], \dots, v[0]$ 其中： $v[0]$ 是 b 序列的低端， $v[L-1]$ 是 b 序列的高端； b 序列长度= $H \times W \times L$ 。

上述的映射过程确保了密码数据 d 到二元序列 b 映射的唯一性。

因此，定义一个密码数据 d 的描述形式为： $d(H, W, L)$ ，当 H 为 8 时，每个基本单元就是一个 `char` 型数据，而当 H 为 32 时，每个基本单元 `int` 型的数据。 H 的大小不受约束，从而可以非常简便地表示任意规模的数据。同时节拍以及节拍序列的划分就可以表示密码数据的分段特征和序列特征。

利用这种逻辑划分方法，可以对密码数据进行整体意义上的编码解释。密码算法中最基本的数据形式是不同规模的比特向量，比特向量以不同的方式组织在一起（如二维比特矩阵）形成一个节拍，例如 AES 算法的一个节拍由 128bit 组成。如果要加密任意规模的数据，只需要在由节拍构成的数据序列上迭代运行。DES 算法的初始数据为 64bit，同时被分为左右 32bit 的数据。因此，DES 的初始数据 `Data` 定义为 `Data (32, 2, 1)`。

另一种常见的数据形式是代替表，简称 `S` 盒，其基本元素是规模相同的基本单元，每一行是由基本单元构成的节拍，而 `S` 盒则是由节拍构成的序列组成。例如，DES 的一个 `S` 盒数据可定义为 `S_Box (4, 16, 4)`，表明该 `S` 盒的由 4 行、16 列的 4bit 的数据元素组成。

采用 `CipherD` 的数据表示模型，可以规范密码数据的表示形式，统一数据的内在结构。与使用通用的整型数，实型数、布尔型数及其数组来描述密码数据相比较，`CipherD` 的表示方法简洁、明了，能准确表达密码数据的特征。同时又可以兼容通用数据类型，是对通用数据类型集合的一个扩充。

2) `CipherD` 的存储模型

密码数据的本原表示形式是二元比特序列 b ，因此如何合适地存储和表示二元序列 b 是实现 `CipherD` 的关键。针对二元序列 b 不同层次的表现需求，本文设计了序列松散型、序列紧凑型 and 单元紧凑型三种数据存储模型，从而以 `CipherD` 的形式为密码数据建立了一个整体框架。

① 序列松散型

密码数据规模不固定的位串运算特征要求能够简单而迅速地定位到二元序列中的每个比特。序列松散型将二元序列的一个比特存储在物理空间的一个字节中。因此，数据规模为 S bit 的密码数据 `Data (H, W, L)` 占用 S 字节的连续存储空间，如图 3.1 所示。

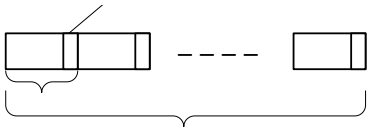


图 3.1 序列松散型

其中，二元序列的高位处于低地址，低位处于高地址，同时将表示二元序列的有效比特置于字节内部的低有效位，为便于等值转换，字节内其他比特用 0 填充。

② 序列紧凑型

序列紧凑型采用物理空间的 1 比特存储二元序列的 1 比特，并且整体上按照二元序列的高位在低地址，低位在高地址的顺序排列，在存储空间的字节内部，高有效位存储二元序列的高位比特，低有效位存储二元序列的低位比特。同时，表示二元序列的有效比特之间紧凑排列，冗余的空间只能出现在存储空间的最后一字节，并且用 0 填充。如图 3.2 所示。

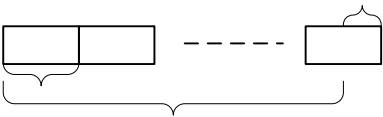


图 3.2 整体紧凑型

因此，数据规模为 S_{bit} 的密码数据 $Data(H, W, L)$ 将占用 $\lceil S/8 \rceil$ 字节的连续存储空间。

③ 单元紧凑型

外在表现上，基本单元通常作为一个数值常量的形式呈现，但其数值常量的规模可以是任意大小，不受底层机器字规模的限制。同时，节拍以及序列的规模也是建立在基本单元之上。因此，基本单元在 CipherD 中占有重要的地位。

单元紧凑型将每一个基本单元作为一个整体考虑，为其分配整数个字节的存储空间，用物理空间的 1bit 存储二元序列的 1bit，有效比特之间紧凑排列。但是，存储一个基本单元时允许有空间的冗余，基本单元与基本单元之间紧密排列。其中，二元序列的高位在低地址，低位在高地址。在存储空间的字节内部，高有效位存储二元序列的高位比特，低有效位存储二元序列的低位比特，冗余的空间用 0 填充。如图 3.3 所示。

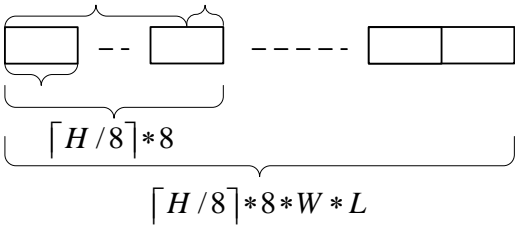


图 3.3 单元紧凑型

因此，密码数据 $Data(H, W, L)$ 将占有 $\lceil H/8 \rceil \times W \times L$ 个字节的连续存储空间。

4 CipherD的实现性能分析

为了使 CipherD 的性能达到最佳，本文采取如下方法：基于 CipherD 的数据表示模型，依据不同存储模型规则分别实现典型的数据抽取操作，然后针对大规模样本数据进行运算，得出操作的时间，并计算出实际存储各种数据时的空间利用率，进而分析比较三种存储模型的性能优劣，为实现最佳性能的 CipherD 提供依据。

在测试中，统一选取了数据高度递增但数据总规模相等（210000bit）的多个数据，测试的环境为：普通的 PC 机（P4 3.2G，512M 内存），Visual C++ 6.0。通过设定抽取参数表，每次抽取相等规模的数据（52500 bit）。将基于三种模型实现数据抽取操作分别连续执行 10000 次，并计算时间消耗。结果表明，基于序列松散型的速度最快，序列紧凑型次之，单元紧凑型最慢，如图 4.1 所示。

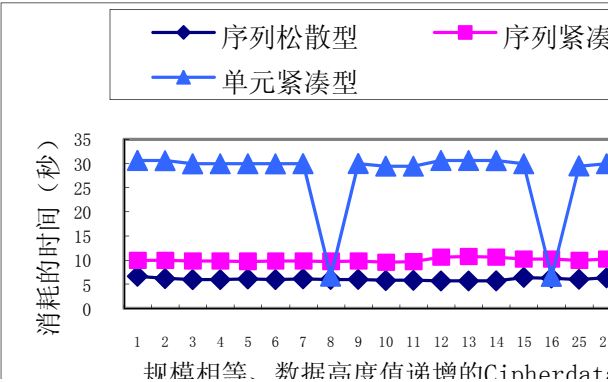


图 4.1 数据抽取的时间

因为在此情形下，需要首先定位到具体的比特，基于序列松散型的实现最为简便，优势明显。

根据每种存储模型的空间分配规则，容易计算存储每个数据时的空间利用率。表 4.1 列举了部分测试数据的情形。

表 4.1 存储数据时的空间利用率

模 型 数 据	序 列 松散型	序 列 紧凑型	单 元 紧凑型
Data (1,300,700)	12.5%	100%	12.5%
Data (3,100,700)	12.5%	100%	37.5%
Data (7,100,300)	12.5%	100%	87.5%
Data (12,100,175)	12.5%	100%	75%
Data (16,125,105)	12.5%	100%	100%

序列松散型由于每个字节内部有 7 比特的冗余，因此空间利用率最低；

当数据规模 S 不是 8 的整数倍时，序列紧凑型仅在存储空间的一个字节冗余 $(8 - S \% 8)$ 个比特，而前 $\lfloor S/8 \rfloor$ 个字节中都是有效比特，因此，空间利用率最高；

对于数据高度 H 不为 8 的整数倍的密码数据，单元紧凑型在每个基本单元所占存储空间的一个字节会有 $(8 - H \% 8)$ 比特的冗余，随着数据高度 H 的增加，空间利用率总体呈现波动上升的趋势，其中，当数据高度为 8 的整数倍时达到最高。

综上所述，序列松散型的性能表现一直稳定，在不关注空间利用率较低的前提下，性能表现最优，基于该存储模型的操作实现简单。序列紧凑型性能表现也较为稳定，在综合考虑时间和空间利用率时，性能能够达到最佳，不过基于该存储模型的实现较为繁琐。单元紧凑型性能表现波动较大，且总体表现不如前两种存储模型，不适用于比特层次的操作选用。

5 小结

本文针对密码数据特有的逻辑层次结构设计了密码数据的层次结构模型 CipherD，并为实

现该模型提出了三种数据存储模型，通过大量的样本数据分析后得出，序列松散型存储在实现 CipherD 时具有较大的优势。随着基于 CipherD 的数据操作研究的深入，可根据具体的需求综合三种存储模型的优劣进行合理的选择，从而全面提高 CipherD 表示密码数据时的整体性能。模型 CipherD 的实现为密码数据确定了一种表示和交换机制，并进一步为密码算法的标准化描述奠定了基础。

参 考 文 献

[1] 吴世忠 陈晓桦 李赫田等. 信息安全测评认证理论与实践[M]. 中国科技大学出版社, 2006.

[2] 谭浩强. C 程序设计[M]. 清华大学出版社, 2002.

[3] Jeffrey R. Lewis, Brad Martin. CRYPTOL: High Assurance, Retargetable Crypto Development and Validation[Z]. MILCOM 2003. IEEE, 2003.

[4] 葛琴. 分组密码算法专用描述语言的研究与实现[D]. 西安: 西安电子科技大学, 2009.

[5] B Schneier. 应用密码学[M]. 机械工业出版社, 2004.

作者简介

周修义 (1985-), 男, 安徽明光人, 硕士研究生, 研究方向为信息安全;

何新民, 男, (1952-), 高级工程师, 硕士生导师, 研究方向为信息安全; 徐莉伟, 女, 工程师, 研究方向为信息安全。